**UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA**

FAIR LINES AMERICA FOUNDATION, INC.,

Plaintiff,

v.

UNITED STATES DEPARTMENT OF COMMERCE and UNITED STATES BUREAU OF THE CENSUS,

Defendants.

Case No. 1:21-cv-1361-ABJ

PLAINTIFF'S MOTION FOR PRELIMINARY INJUNCTION

Fair Lines America Foundation, Inc. ("Plaintiff") moves for a Preliminary Injunction against the United States Department of Commerce and the United States Bureau of the Census (collectively "Defendants") to enjoin Defendants from failing to comply with Plaintiff's March 31, 2021 Freedom of Information Act request ("the Request") by withholding non-exempt records under FOIA, 5 U.S.C. § 552, and to order Defendants to produce all responsive non-exempt records and data improperly withheld from the May 25 production within 10 days of the date of the Court's Order, or before August 15, 2021, whichever is earlier. Plaintiff also respectfully requests an order from this Court that Defendants produce all non-exempt responsive records and data from Defendants' identified potentially responsive emails (that have not yet been produced to Plaintiff) as soon as practicable, and order Defendants to produce a *Vaughn* Index specifically describing in detail each record and portion thereof withheld as exempt within the same timeframe.

The present and continuing harms brought about by Defendants' inaction, along with the increasing threat of irreparable harm from their continued delay in light of the impending release of the Census Bureau's "legacy format" data, the redistricting cycle that will immediately follow, and the approaching elections, necessitates this request for preliminary injunctive relief.

This motion is made on the grounds specified in this motion, Plaintiff's brief in support thereof (along with supporting Exhibits), the Complaint (along with supporting Exhibits), and such other and further evidence as may be presented to the Court.

Dated: July 19, 2021

Respectfully submitted,

/s/ Jason Torchinsky
Jason Torchinsky (D.C. Bar No. 976033)
jtorchinsky@holtzmanvogel.com
Jonathan P. Lienhard (D.C. Bar No. 474253)
jlienhard@holtzmanvogel.com
Kenneth C. Daines (D.C. Bar No. 1600753)
kdaines@holtzmanvogel.com
HOLTZMAN VOGEL BARAN TORCHINSKY &
JOSEFIAK PLLC
15405 John Marshall Highway
Haymarket, VA 20169
Phone: (540) 341-8808
Fax: (540) 341-8809
Counsel for Plaintiff

CERTIFICATE OF SERVICE

I do hereby certify that, on this 19th day of July 2021, the foregoing Application for Preliminary Injunction was filed electronically with the Clerk of Court using the CM/ECF system. The system instantaneously generated a Notice of Electronic Filing which served all counsel of record.

/s/ Jason Torchinsky____________
Jason Torchinsky (D.C. Bar No. 976033)
jtorchinsky@holtzmanvogel.com
Jonathan P. Lienhard (D.C. Bar No. 474253)
jlienhard@holtzmanvogel.com
Kenneth C. Daines (D.C. Bar No. 1600753)
kdaines@holtzmanvogel.com
HOLTZMAN VOGEL BARAN TORCHINSKY & JOSEFIAK PLLC
15405 John Marshall Highway
Haymarket, VA 20169
Phone: (540) 341-8808
Fax: (540) 341-8809
Counsel for Plaintiff

UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA

FAIR LINES AMERICA FOUNDATION, INC., Plaintiff, v. UNITED STATES DEPARTMENT OF COMMERCE and UNITED STATES BUREAU OF THE CENSUS, Defendants.	Case No. 1:21-cv-1361-ABJ **PLAINTIFF'S STATEMENT OF POINTS AND AUTHORITIES IN SUPPORT OF MOTION FOR PRELIMINARY INJUNCTION**

The Constitution requires that the federal government conduct an "actual Enumeration" of the population of the United States every ten years—known as the decennial Census—to determine the total population of each state and "apportion" the seats in the House of Representatives between the states. U.S. Const. art. I, § 2, cl. 3. In the late 1990s, the Census Bureau proposed using statistical methods to "adjust" census numbers used in the apportionment using various statistical methods. The Supreme Court rejected this method because the Census Act prohibited the proposed uses of statistical sampling in calculating the population for purposes of apportionment. *Dep't of Com. v. U.S. House of Representatives*, 525 U.S. 316, 334 (1999); *see also id.* at 349 (Scalia, J., concurring) ("[A]n apportionment census conducted with the use of 'sampling techniques' is not the 'actual Enumeration' that the Constitution requires."). Several years later, Utah challenged the Census Bureau's use of "household imputation" to fill in data on certain missing households—essentially by borrowing data from a nearby neighbor and "imputing" that information to the missing household. In a divided opinion, the Supreme Court approved this use of household imputation as not inconsistent with the Constitution's "actual Enumeration" requirement and determined it was not a prohibited use of statistical sampling. *Utah v. Evans*, 536 U.S. 452, 457 (2002).

For the 2020 Census, the Census Bureau is now for the first time ever using a methodology it has termed "group quarters imputation" to fill in apparently missing or incomplete data for certain group housing facilities—ranging from jails, prisons, nursing homes, military bases, to colleges and universities. This methodology was not publicly revealed through any sort of notice and comment process, and to this day the methodology, nature, and extent of how many people were added (and in what states) through "group quarters imputation" remains undisclosed.

In response to the FOIA request that is the subject of this case, the Census Bureau revealed some information about their various attempts to test different methods of "group quarters imputation." For the first set of documents the Census Bureau produced, however, certain potentially crucial information about whether the Census Bureau used statistical methods in ascertaining the actual enumeration was withheld by the Census Bureau under its newly claimed interpretation of Title 13's privacy protections.

Plaintiff Fair Lines America Foundation ("Plaintiff" or "Fair Lines") is seeking information to help inform the public about the nature and extent of this new methodology. As a result, Plaintiff respectfully requests that this Court issue a preliminary injunction to enjoin Defendants the United States Department of Commerce and the United States Census Bureau ("Defendants") from continuing to violate FOIA's requirements by improperly redacting and withholding non-exempt records sought in Plaintiff's March 31, 2021 FOIA request (the "Request"). Because Defendants initially failed to communicate their determination as to whether they will comply with Plaintiff's Request within FOIA's applicable statutory timeframe prior to filing suit, Plaintiff constructively exhausted its administrative remedies, and this Court now has jurisdiction to afford all necessary relief to Plaintiff to ensure full compliance with Plaintiff's Request.

Under these unique circumstances, with a high-stakes and time-sensitive matter like the 2020 Census, the Census Bureau's publicly announced irregularities in imputation of its group quarters data due to the COVID-19 pandemic, and the Bureau's August 16 redistricting data release deadline, time is truly of the essence—preliminary injunctive relief and the immediate public release of the requested information is necessary to avoid irreparable harm to Plaintiff, and indeed the public at large. Because Plaintiff is likely to succeed on the merits of its claims that Defendants' interpretation of Title 13's confidentiality requirements is plainly erroneous (which is a pure question of law), has demonstrated a likelihood of irreparable harm, and the balance of the equities and public interest factors favor relief for Plaintiff, this Court should grant Plaintiff's motion for a preliminary injunction. Accordingly, Plaintiff respectfully requests that this Court order that Defendants disclose all non-exempt withheld information and data responsive to Plaintiff's Request within 10 days of the Court's order (or no later than August 15, 2021) to avoid irreparable harm to Plaintiff and the American public. The disclosure requested here will likely either put to rest concerns about the Bureau's new methodology, or become evidence needed to prevent the use of improperly imputed apportionment data. If the Census Bureau is permitted to conduct these sorts of methodology changes and implementations behind closed doors and without the sunlight that FOIA and Title 13 require, electoral chaos may result from the states' reliance on potentially defective numbers in conducting redistricting.

STATEMENT OF FACTS

As the Constitution mandates, a census must be conducted every ten years "in such Manner as [Congress] shall by Law direct" to reapportion the number of seats allocated to each state in the House of Representatives. U.S. Const., art. I, § 2, cl. 3. The state population totals are also used "to allocate federal funds to the States and to draw electoral districts." *Dep't of Com. v. New York*, 139 S. Ct. 2551, 2561 (2019). Congress has delegated the taking of the census to the Secretary of Commerce "in such form and content as he may determine," 13 U.S.C. § 141(a), with the Census Bureau being the entity primarily responsible for administering the same.

Due largely to challenges stemming from the global COVID-19 pandemic, administration of the 2020 Census has been anything but smooth. For instance, in late 2020, well after Census Day had passed, public reporting described "processing anomalies" of census records for the 2020 national tally that "if left unfixed, could miscount millions of people."[1] "[M]ajor inconsistencies" unearthed by the Census Bureau largely centered around "the information it has gathered this year about residents of college dorms, prisons and other group living quarters—a category that, for the 2020 census, included around 8 million people."[2]

Consequently, on February 12, 2021, the Census Bureau publicly announced that the first release of its redistricting data, which was originally scheduled to be delivered to the states by March 31, 2021, would be delayed until September 30, 2021.[3] On March 15, 2021, following lawsuits filed by the State of Ohio and the State of Alabama, the Census Bureau announced that there would be a public release of the "legacy format" summary redistricting data (which states

[1] Hansi Lo Wang, *Millions of Census Records May Be Flawed, Jeopardizing Trump's Bid to Alter Count*, NPR (December 5, 2020), https://www.npr.org/2020/12/05/943416487/millions-of-census-records-may-be-flawed-jeopardizing-trumps-bid-to-alter-count (accessed on July 18, 2021) [hereinafter *Millions of Census Records May Be Flawed*]; *see also* Wang, *6-Month Delay in Census Redistricting Data Could Throw Elections Into Chaos*, NPR (February 12, 2021), https://www.npr.org/2021/02/12/965823150/6-month-delay-in-census-redistricting-data-could-throw-elections-into-chaos (accessed on July 18, 2021) [hereinafter *6-Month Delay in Census Redistricting Data Could Throw Elections Into Chaos*].

[2] *Millions of Census Records May Be Flawed, supra.*

[3] Press Release, *Census Bureau Statement on Redistricting Data Timeline*, U.S. Census Bureau (Feb. 12, 2021), https://www.census.gov/newsroom/press-releases/2021/statement-redistricting-data-timeline.html (accessed on July 18, 2021) [hereinafter *Feb. 12, 2021 Census Press Release*].

are assured they can rely on for accuracy in conducting redistricting) on August 16, 2021.[4] The Bureau explained that its delays were necessary largely to allow for time to address difficulties and irregularities it encountered while gathering and tabulating group quarters data for the 2020 Census due to the COVID-19 pandemic.[5] Specifically, the Bureau's Chief of Decennial Statistical Studies Division, acknowledged on the Bureau's public website that it "had to adapt and delay some of the ways we counted group quarters because of the COVID-19 pandemic," and that, consequently, "[a]fter the end of data collection, when we began processing census data from group quarters, we realized that many of them were occupied on April 1, 2020 (the reference day for the census), but didn't provide a population count."[6] The Bureau also explained the significant impact such group quarters data discrepancies can have for obtaining an accurate population count:

> [W]hen we enumerated [group quarters] in midsummer, some group quarters said they were vacant but they were actually occupied on April 1. If not corrected, such cases could lead to an undercount. If the corrections were not properly coordinated with our procedures to remove duplicated people, they could contribute to an overcount.[7]

Accordingly, the Bureau announced that it is now using a new "group quarters count imputation"

[4] Press Release, *U.S. Census Bureau Statement on Release of Legacy Format Summary Redistricting Data File*, U.S. Census Bureau (Mar. 15, 2021) https://www.census.gov/newsroom/press-releases/2021/statement-legacy-format-redistricting.html (accessed on July 16, 2021); Important Dates, U.S. Census Bureau, https://2020census.gov/en/important-dates.html (accessed on May 21, 2021).

[5] *Feb. 12, 2021 Census Press Release*, *supra*; *see also* Press Release, *Census Bureau Statement on Modifying 2020 Census Operations to Make Sure College Students Are Counted*, U.S. Census Bureau (Mar. 15, 2020), https://www.census.gov/newsroom/press-releases/2020/modifying-2020-operations-for-counting-college-students.html (accessed July 18, 2021) [hereinafter *Mar. 15, 2020 Census Press Release*]; Press Release, *2020 Census Operational Adjustments Due to COVID-19*, U.S. Census Bureau, https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/operational-adjustments.html (accessed on July 18, 2021) [hereinafter *Operational Adjustments*].

[6] Pat Cantwell, *How We Complete the Census When Households or Group Quarters Don't Respond*, U.S. Census Bureau (April 16, 2021), https://www.census.gov/newsroom/blogs/random-samplings/2021/04/imputation-when-households-or-group-quarters-dont-respond.html (accessed on July 18, 2021) [hereinafter *Pat Cantwell Statement*].

[7] *Id.*

procedures on unresolved group quarters, including for missing characteristics such as age or race, even though that method "had never before [been] conducted" for group quarters previously.[8] As a result, tabulation and verification of the final results of the 2020 Census remains ongoing,[9] while public uncertainty about reliability of the data remains high and questions about imputation method(s) have been largely unaddressed by the Bureau.

It is regarding this 2020 Census group quarters data that Plaintiff[10] submitted a FOIA request on February 19, 2021. Ex. 1 ("Declaration of Adam Kincaid") at ¶ 5. Specifically, Plaintiff requested records demonstrating or reflecting the number of residents reported by housing facilities nationwide in response to the Census Bureau's 2020 Group Quarters Enumeration questionnaire and information about the methodology. *See* Compl. Ex. A, ECF No. 1-1. The Census Bureau denied Plaintiff's request on March 12, 2021, asserting that the requested records were exempt from disclosure under Section 9 of the Census Act. Compl. Ex. B, ECF No. 1-2.

In response to this denial, Plaintiff submitted a new FOIA request on March 31, 2021 ("the Request"). The Request clarifies that Plaintiff only seeks summaries, tabulations, and other statistical materials derived from, summarizing, or otherwise relating to the original underlying group quarters population data reported for the 2020 Census, rather than the underlying raw data itself and the methodology. *See* Compl. Ex. C, ECF No. 1-3 at 3-4. Specifically, Plaintiff stated that it does not "seek disclosure of the underlying raw group quarters population data itself as originally 'reported by, or on behalf of, any particular respondent' to the Bureau, 13 U.S.C. § 8(b),"

[8] *Id.*

[9] *Important Dates*, *supra*.

[10] Plaintiff Fair Lines is a Section 501(c)(3) non-profit organization interested in openness and transparency in government, with an emphasis on educating the public and ensuring fair and legal enumeration, apportionment, and redistricting processes. To that end, Fair Lines reviews and publicizes records in the possession of Defendants in light of the Census Bureau's public announcements of its difficulties and various concerns regarding the gathering and counting of group quarters data for the 2020 Census. *See* Ex. 1 ("Declaration of Adam Kincaid") at ¶ 4.

nor "any 'publication whereby the data furnished by any particular establishment or individual under this title can be identified,' 13 U.S.C. § 9(a)(2)." *Id.*; *see also* Ex. 1 at ¶ 6.

Plaintiff also requested expedited processing of the Request based on its compelling need for the records and the urgency of informing the public of any irregularities in 2020 Census data given the time-sensitive nature of the redistricting process leading up to the impending election season, as well as the decennial nature of the Census Bureau's data collection. Compl. Ex. C, ECF No. 1-3 at 6-7. Finally, Plaintiff requested a fee waiver or limitation of fees because the records are likely to contribute significantly to public understanding of the operations of the Government and is for non-commercial purposes. *Id.* at 5-6.

On April 7, 2021, having received no confirmation that the Request was received by the Census Bureau, Plaintiff, through its counsel, sent an email to the Census Bureau inquiring about the status of the Request. *See* Compl. Ex. D, ECF No. 1-4. Plaintiff received two automated messages in response, eventually assigning the Request tracking number DOC-CEN-2021-001311. *See* Compl. Ex. E, ECF No. 1-5. On April 13, 2021, the Census Bureau's FOIA Section Chief affirmed that the Request had been received and that a search had commenced. *See* Compl. Ex. F, ECF No. 1-6 at 1.[11] However, after the FOIA statutory twenty-business-day deadline

[11] Confusingly, the Census Bureau created two different tracking numbers for this single Request, causing Plaintiff to receive two separate automated messages on April 7, 2021, containing the two tracking numbers. Compl. Ex. E, ECF No. 1-5. Additionally, the automated messages both indicated that the Request had been "submitted" on April 7, 2021, even though the request was submitted on March 31, 2021, to the Census Bureau's designated email address for submitting FOIA requests, Census.efoia@census.gov. Compl. Ex. F, ECF No. 1-6 at 3. Plaintiff's counsel sent an email on April 8, 2021, inquiring about these discrepancies, but did not receive a response. Then on April 12, 2021, Plaintiff received another automated message saying that one of the requests had been "processed with the following final disposition: Duplicate Request." On April 12, Plaintiff's counsel again emailed the Bureau to ask about the status of its Request and to follow up on its unanswered questions from the April 8 email, to which the Bureau on April 13, 2021 only provided a partial answer that the request was evidently forwarded to DOC, but was then closed as a duplicate request because the Census Bureau determined it was better suited to process the Request. *See id.* at 1-2.

(calculated from the date Fair Lines emailed the Request to the Census Bureau) passed on April 28, 2021, *see* 5 U.S.C. § 552(a)(6)(A)(i), Fair Lines had still received no determination from Defendants regarding the Request, including no decision on its application for expedited processing. Ex. 1 at ¶¶ 8-9.

After the April 13, 2021 email from the Census Bureau, Plaintiff received no further communications from the Census Bureau until it filed its Complaint with this Court on May 18, 2021, s*ee id.* ¶ 9; Compl. ECF No. 1, having constructively exhausted all administrative remedies. Soon after the Complaint was filed, the Bureau's FOIA Analyst sent Plaintiff's counsel an email notifying Plaintiff simply that the Bureau is "diligently working on your FOIA request." Ex. 2 ("Census Bureau's Post-Complaint Email Correspondence to Plaintiff"). Then, on May 19, 2021, this same analyst wrote another email to Plaintiff's counsel, repeating that the Bureau is "diligently working on your FOIA request" but added that "in order to conduct an email search for this request, we will need a date range for the emails to search." *Id.*

On May 25, 2021, Defendants sent a letter (dated May 24, 2021) to Plaintiff's counsel partially granting and partially denying Plaintiff's FOIA request, providing Plaintiff with 988 pages of redacted responsive records. Ex. 4 ("May 24 Census Bureau Determination Letter and Production"). Of those, 166 pages were either fully or partially redacted. *See id.* No records from 2021 were included in the production; *i.e.*, all produced (visible) records were dated December 2020 and earlier. *See id.* Defendants claimed all withheld portions were redacted "pursuant to FOIA Exemptions 3 and 5, Title 5, United States Code, Sections 552(b)(3) and (b)(5)." *Id.* Of greatest relevance to this action, Defendants asserted that information withheld under Exemption 3 is "protected by Title 13, United States Code, Section 9," which Defendants interpret to mean "requires that census records be used solely for statistical purposes and makes these records confidential." *Id.*

Counsel for both parties met the following day in a telephonic consultation, with Plaintiff's counsel requesting that Defendants (1) review the May 25 production to clarify which exemption applied to each redaction, (2) produce all post-December 2020 responsive records, and (3) produce the responsive emails referenced in the May 19 correspondence. In a follow-up email on May 26, Plaintiff's counsel agreed to narrow the scope of the unresolved email search to "all responsive emails sent or received between March 31, 2020 and March 31, 2021," *See* Ex. 3 ("Correspondence Between Parties' Counsel") at 18-19 (5.26.21 Kossak email), and reiterated his client is seeking "only *aggregated* numbers on a statewide or county-wide level" that were counted as a result of group quarters imputation procedures. Ex. 3 at 18-19 (5.26.21 Torchinsky email). Plaintiff's counsel also clarified that Plaintiff was not requesting any exempt "underlying raw group quarters population data as originally 'reported by, or on behalf of, any particular respondent' to the Bureau," *id.* (quoting 13 U.S.C. § 8(b)), nor was Plaintiff seeking any "'publication whereby the data furnished by any particular establishment or individual under this title can be identified,'" *id.* (quoting 13 U.S.C. § 9(a)(2)), or other "'individual reports,'" *id.* (quoting 13 U.S.C. § 9(a)(3)). Both parties' counsel also discussed the remaining, but not yet produced, responsive records, all of which were created after December 2020. Ex. 3 at 16 (6.16.21 Torchinsky email).

On May 27, 2021, Defendants' counsel conveyed that his client had agreed to review the May 25 production to "determine whether they stand by those redactions" and to clarify the basis for each redaction. Ex. 3 at 18 (5.27.21 Kossak email). Defendants' counsel was unable to provide a timetable at that point for completing this process. *Id.* Plaintiff's counsel responded with a request to receive additional documents on a rolling basis as they were ready for release, to which Defendants' counsel did not reply. Ex. 3 at 18 (5.27.21 Torchinsky email). The following day, the

Census Bureau granted Plaintiff's requests for expedited processing of the FOIA request and for a fee waiver. Ex. 5 ("May 27 Correspondence Granting Expedited Processing").

On June 8, 2021, because Plaintiff had not heard from Defendants or received either the reprocessed May 25 production or any of the requested emails, Plaintiff's counsel emailed Defendants' counsel requesting an update. Ex. 3 at 17 (6.8.21 Torchinsky email). Defendants' counsel did not have an answer at that time, and eventually responded over a week later on June 16, 2021, stating that Defendants would release the re-processed May 25 production to Plaintiff by June 24, 2021. Ex. 3 at 16 (6.16.21 Kossak email). Defendants' counsel did not provide any information on the status of the requested emails at that time. *Id.* On June 21, 2021, Defendants filed an answer to Plaintiff's complaint. ECF No. 7.

On June 24, Plaintiff's counsel emailed Defendants' counsel to ask when the reprocessed records would be released and if Defendants had provided any answers on the additional emails and post-December 2020 records. Ex. 3 at 14 (6.24.21 Torchinsky email). Defendants' counsel responded that the re-processed records would not be provided to Plaintiff by the promised June 24 deadline because Defendants claimed to have run into "unexpected technical difficulties" and that they hoped "to have the document available by the end of [June]." Ex. 3 at 14 (6.24.21 Kossak email). In response, Plaintiff's counsel explained that given the potentially time sensitive nature of the information contained in these records, and because the parties' agreement had been unilaterally pushed back by Defendants, he would consult with his client about seeking a preliminary injunction regarding the withheld records. Ex. 3 at 13 (6.25.21 Torchinsky email).

In a June 25 email, Defendants' counsel provided specific pages corresponding with particular justifications for the redactions from the May 25 production. Ex. 3 at 11-13 (6.25.21 Kossak email). Defendants stood by all of their redactions, and asserted that the majority of the information withheld was redacted "to ensure that every information product released by the Census Bureau adheres to the confidentiality requirement of Title 13 and other applicable statutes," making that information all allegedly exempt from disclosure under FOIA Exemption 3. *Id.* The email also indicated Defendants had found 2,600 emails that were potentially responsive to the Request, and that Defendants would agree in the Joint Status Report due on July 20, 2021 to "using their best efforts to process 300 pages of potentially responsive records per month, with the first release of any nonexempt, responsive records by July 30, 2021." *Id.*

Having finally received explanations for the redactions in the May 25 production, and after a Zoom call between parties' counsel on June 29, 2021, Plaintiff's counsel provided Defendants' counsel with a list of redactions Plaintiff views to be improper along with an attached excerpt of those pages, with the most glaring issues arising from withholdings of summary statistical information and tabulations that Plaintiff indicated are subject to disclosure under 13 U.S.C. § 8(b). *See* Ex. 7 ("June 29 Email—Plaintiff's Challenged Redactions"). Plaintiff's counsel also requested an update on the status of the search for the responsive emails and post-December 2020 records. *Id.*

Defendants' counsel responded in a July 6, 2021 email, providing Plaintiff with just two responsive post-December 2020 records. Ex. 6 ("July 6 Additional Production"). In the same correspondence, Defendants' counsel again defended all of the redactions in the May 25 production, asserting that because of the "risk of re-identification attacks on aggregated data releases" in the modern age of computing power and sophistication, Defendants "generally avoid[] the release of intermediate work product because it can be used in combination with other intermediate work products, official publications, and the final product to re-identify individual respondents and their data items"; accordingly, Defendants maintain that release of *any* of the aggregate or summary data withheld from Plaintiff would violate Title 13's confidentiality provisions. *See id.* Defendants' counsel asserted that Plaintiff "[has] not identified any particular reason why the redacted data is needed urgently," even though Defendants had previously granted Plaintiff's request for expedited processing on May 28, 2021. *Id.*; Ex. 3 at 17 (5.28.21 Kossak email). Finally, Defendants' counsel indicated that Defendants had identified 917 potentially responsive emails (in contrast with the "2,600 potentially responsive emails" mentioned in Defendants' June 25 email, *see* Ex. 3 at 11-12 (6.25.21 Kossak email)) consisting of 25,899 pages of material, and reaffirmed Defendants' initial offer to attempt review of 300 pages of emails per month for potential release to Plaintiff. Ex. 3 at 6-9 (7.6.21 Kossak email).[12]

On July 10, Plaintiff's counsel responded that because of the significant and time sensitive nature of Plaintiff's request, it would be seeking a preliminary injunction seeking production of the improperly withheld/redacted, non-exempt pages of the May 25 production, particularly in light of the Census Bureau's impending August 16 release of the legacy format summary data and the redistricting process that will commence in earnest immediately afterward. Ex. 3 at 6 (7.10.21 Torchinsky email). Plaintiff's counsel also proposed substantially narrowing the universe of remaining emails for searching to focus on those most urgently sought by Plaintiff, namely imputed statewide group quarters population totals, while excluding any county- or local-level numbers or tabulations, and requested an estimated production timeline under these proposed parameters. *Id.*

To date, Plaintiff has not received a single email responsive to the Request (submitted on March 31, 2021), nor has it received any of the improperly withheld pages or redacted information from the May 25 production, which prevents Plaintiff's access to information not exempt from

[12] At the production rate proposed by the Census Bureau, it would take more than 7 years for the Plaintiff to receive all of the responsive records.

disclosure under Title 13.[13]

Plaintiff here challenges only certain of the Defendants' redactions. Certain redactions, such as descriptions of internal computer file locations or information that appears to discuss only a single institution's group quarters, are not being challenged. Plaintiff attaches hereto Exhibit 7, which is an excerpt of the 988-page production containing the redacted and withheld pages Plaintiff is challenging, along with the June 29, 2021 email from Plaintiff's counsel to Defendants' counsel that provides a narrative description of many of the redacted pages along with Plaintiff's explanation of why each appears to be an improper redaction. Ex. 7.

STANDARD OF REVIEW

In addition to this Court's equitable authority to enjoin and order compliance with FOIA, FOIA itself provides a reviewing court authority "to enjoin the agency from withholding agency records and to order the production of any agency records improperly withheld from the complainant." 5 U.S.C. § 552(a)(4)(B). Plaintiff requests that this Court issue a preliminary injunction prohibiting the Defendants from continuing to redact or withhold the requested information and documents where they are not subject to a proper FOIA exemption. [14]

[13] Defendants' frequent delays and dawdling in producing responsive documents, along with its current proposed timeline for producing the outstanding responsive emails, also run contrary to FOIA's statutory demands, especially given the pressing and time-sensitive nature of the Request and the fact that expedited processing was granted for Plaintiff's FOIA request.

[14] Separately, Plaintiff requests that the Court order that Defendants produce a *Vaughn* index describing each document claimed as exempt with sufficient specificity "to permit a reasoned judgment as to whether the material is actually exempt under FOIA." *Founding Church of Scientology, Inc. v. Bell*, 603 F.2d 945, 949 (D.C. Cir. 1979); *King v. U.S. Dep't of Justice*, 830 F.2d 210, 223–24 (D.C. Cir. 1987) (a *Vaughn* index must "describe each document or portion thereof withheld, and for each withholding it must discuss the consequences of disclosing the sought-after information."). Additionally, Plaintiff requests that the Court order Defendants to disclose any "reasonably segregable" non-exempt portions of the fully redacted pages as required by FOIA, 5 U.S.C. § 552(b), and that if Defendants assert that a record contains non-exempt segments that are so dispersed throughout the records as to make segregation impossible, Defendants must still indicate what portion of the document is non-exempt, and describe how the

The preliminary injunction standard is well understood in this Court. This Court recently explained that "[a] preliminary injunction is 'an extraordinary remedy that may only be awarded upon a clear showing that the plaintiff is entitled to such relief.'" *Elec. Privacy Info. Ctr. v. DOJ*, 15 F. Supp. 3d 32, 38 (D.D.C. 2014) (citing *Winter v. Nat. Res. Def. Council, Inc.*, 555 U.S. 7, 22 (2008)). To obtain a preliminary injunction, the moving party must show: "(1) a substantial likelihood of success on the merits; (2) that it would suffer irreparable injury if the injunction is not granted; (3) that the balance of equities tips in its favor; and (4) that the public interest would be furthered by the injunction." *Wash. Metro. Area Transit Auth. v. Local 689, Amalgamated Transit Union*, 113 F. Supp. 3d 121, 126 (D.D.C. 2015) (citing *Winter*, 555 U.S. at 20); *see also Coalition for Parity, Inc. v. Sebelius*, 709 F. Supp. 2d 6, 7-8 (D.D.C. 2010); *Hall v. Johnson*, 599 F. Supp. 2d 1, 6 n. 2 (D.D.C. 2009). Further, the balance of the equities and public interest preliminary injunction factors "merge when the Government is the opposing party.'" *Nken v. Holder*, 556 U.S. 418, 435 (2009).

"In conducting an inquiry into these four factors, [a] district court must balance the strengths of the requesting party's arguments in each of the four required areas." *Elec. Privacy Info. Ctr.*, 15 F. Supp. 3d at 38 (internal citation and quotation marks omitted). "The District of Columbia Circuit applies a 'sliding-scale' approach to the preliminary injunction factors, meaning that 'a strong showing on one factor could make up for a weaker showing on another.'" *Indian River Cnty. v. Rogoff*, 110 F. Supp. 3d 59, 67 (D.D.C. 2015) (citation omitted). The D.C. Circuit places preeminent importance on the first and second factors, indicating that plaintiffs must "independently show both a likelihood of success on the merits and irreparable harm." *Brennan Ctr. for Justice v. Dep't of Com.*, 498 F. Supp. 3d 87, 96 (D.D.C. 2020). However, Plaintiff's

material is dispersed through the document. *See Mead Data Cent. v. U.S. Dep't of the Air Force*, 566 F.2d 242, 261 (D.C. Cir. 1977).

"probability of success on the merits is the most critical of the criteria when considering a motion for a preliminary injunction." *Carey v. FEC*, 791 F. Supp. 2d 121, 128 (D.D.C. 2011).

Plaintiff satisfies all elements for the issuance of a preliminary injunction as demonstrated in the discussion that follows.

ARGUMENT

I. PLAINTIFF IS ENTITLED TO A PRELIMINARY INJUNCTION.

A. Plaintiff is Likely to Succeed on the Merits of its FOIA Claim.

1. Because Defendants failed to meet their statutory deadline under FOIA, Plaintiff's administrative remedies have been constructively exhausted—this Court has jurisdiction to ensure full compliance.

The Freedom of Information Act requires that each federal agency, upon receiving any reasonably articulated request that accords with the agency's published rules, "shall make the records promptly available to any person." 5 U.S.C. § 552(a)(3)(A). The agency has twenty business days from the date it receives the request to determine "whether to comply with such request" and to "immediately notify the person making such request" of "such determination and the reasons therefor" and of the "right of such person to appeal to the head of the agency" any adverse determination. *Id.* § 552(a)(6)(A)(i). In "unusual circumstances," the agency can provide written notice to the requestor setting forth the circumstances warranting an extension of time, and can extend this response period for no "more than ten working days." *Id.* § 552(a)(6)(B)(i). Then, FOIA again instructs that responsive, non-exempt records "shall be made promptly available to such person making such request." *Id.* § 552(a)(6)(C)(i).

FOIA's requirement that records be made "promptly available" after an agency's determination "typically [means] within *days or a few weeks* of a 'determination', not months or years." *CREW v. FEC*, 711 F.3d 180, 188-89 (D.C. Cir. 2013) (Kavanaugh, J.) (quoting 5 U.S.C.

§ 552(a)(3)(A), (a)(6)(C)(i)) (emphasis added). "[A]n agency's failure to comply with the FOIA's time limits is, by itself, a violation of FOIA . . . ." *Gilmore v. U.S. Dep't of Energy*, 33 F.Supp.2d 1184, 1187 (N.D. Cal. 1998) (citation omitted). Even if a delay in processing a request results from "bureaucratic mishandling rather than intentional obfuscation" that is not enough on its own to "make the delay reasonable" under the statute. *See Munger, Tolles & Olson LLP v. U.S. Dep't of Army*, 58 F. Supp. 3d 1050, 1056 (C.D. Cal. 2014).

When the agency does not respond by the statutory deadline, the requestor may sue in federal court without exhausting internal agency appeal processes, as Plaintiff did here. 5 U.S.C. § 552(a)(6)(C)(i) ("Any person making a request to any agency for records . . . shall be deemed to have exhausted his administrative remedies with respect to such request if the agency fails to comply with the applicable time limit provisions of this paragraph."); *Nurse v. Sec'y of the Air Force*, 231 F. Supp. 2d 323, 328 (D.D.C. 2002) ("The FOIA is considered a unique statute because it recognizes a constructive exhaustion doctrine for purposes of judicial review upon the expiration of certain relevant FOIA deadlines.").

Because Defendants violated FOIA by failing to communicate their determination within the statutory deadlines, and only produced records after Plaintiff filed its complaint in this Court, administrative remedies have been exhausted, and this Court can maintain jurisdiction to ensure that Defendants properly and completely fulfill Plaintiff's Request. Plaintiff's Request at issue in this litigation was emailed on March 31, 2021 to the Census Bureau's designated email address for receiving FOIA requests, Census.efoia@census.gov, *see* Compl. Ex. F, ECF No. 1-6, and the April 28, 2021 deadline of twenty business days passed without receipt of any notification of Defendants' determination whether to comply with the Request prior to filing the Complaint, 5 U.S.C. § 552(a)(6)(A)(i). Defendants thus violated FOIA's plain statutory deadlines, and the doctrine of constructive exhaustion of administrative remedies permits this Court to exercise jurisdiction over this action to ensure the agency fully complies with all of FOIA's requirements.

2. Because Plaintiff's Request explicitly seeks non-exempt records, Title 13 of the Census Act does not prohibit Defendants from fulfilling Plaintiff's Request.

Congress's intent in enacting FOIA was to implement "a general philosophy of full agency disclosure unless information is exempted under clearly delineated statutory language." *U.S. Dep't of Def. v. Fed. Lab. Rels. Auth.*, 510 U.S. 487, 494 (1994) (quoting *Dep't of Air Force v. Rose*, 425 U.S. 352, 360-61 (1976)). Accordingly, FOIA "creates a strong presumption in favor of disclosure," *Davin v. U.S. Dep't of Justice*, 60 F.3d 1043, 1049 (3d Cir. 1995), requiring "the fullest possible disclosure of an agency's records." *Larson v. Dep't of State*, No. 1:02cv01937 (PLF), 2005 U.S. Dist. LEXIS 35713, at *7 (D.D.C. Aug. 10, 2005). Consistent with FOIA's demanding disclosure requirement, FOIA's nine "narrowly-tailored exemptions," *Larson*, 2005 U.S. Dist. LEXIS 35713, at *7, have been "consistently given a narrow compass," *U.S. Dep't of Justice v. Tax Analysts*, 492 U.S. 136, 151 (1989). The agency "bears the burden of establishing the applicability of the claimed exemption." *Assassination Archives & Research Ctr. v. CIA*, 334 F.3d 55, 57 (D.C. Cir. 2003).

As explained above, the Census Bureau denied Plaintiff's previous February 19, 2021 FOIA request in its entirety, citing 13 U.S.C. § 9 and stating that Title 13 "requires that census records be used solely for statistical purposes and makes these records confidential." Compl. Ex. B, ECF No. 1-2 at 1. In response, Plaintiff filed its subsequent March 31, 2021 Request at issue here only seeking disclosure of data that Title 13 expressly permits. Indeed, Plaintiff's Request states with unmistakable clarity that it does not seek non-exempt data whatsoever: Plaintiff's Request expressly states that it does <u>not</u> "seek disclosure of the underlying raw group quarters population data itself as originally 'reported by, or on behalf of, any particular respondent' to the Bureau, 13 U.S.C. § 8(b), nor do we seek any 'publication whereby the data furnished by any particular establishment or individual under this title can be identified,' 13 U.S.C. § 9(a)(2)." Compl. Ex. C, ECF No. 1-3 at 3. Plaintiff's Request also clarifies that it does not request examination of protected underlying "individual reports," 13 U.S.C. § 9(a)(3), at all. Rather, it only seeks summaries, "tabulations[,] and other statistical materials," 13 U.S.C. § 8(b), "*deriving from* or summarizing the originally reported raw data, and/or records with data that has been *reformulated* or *repurposed* by the Bureau in a form such that the underlying data can no longer be identified with a particular establishment or individual." Compl. Ex. C, ECF No. 1-3 at 3. Because Plaintiff does not request data that is exempt from disclosure under the Census Act, the Census Bureau cannot rely on that Act's exemptions to effectively deny Plaintiff's request through targeted redactions and withheld records.

Nevertheless, in its 988-page May 25 production, the Census Bureau withheld and redacted nearly all of Plaintiff's requested non-exempt summary data interspersed throughout the production, essentially amounting to a constructive denial of Plaintiff's Request. The overwhelming majority of Defendants' redactions and withheld records cite Title 13 as providing a statutory bar to disclosure of all Census data, allegedly making it exempt from production under FOIA Exemption number 3. Such redactions, and in many instances fully withheld records, are tantamount to the Bureau's earlier denial of Plaintiff's February request because the effect is the same: release of the requested statistical information has been denied. *Cf. Thomas v. HHS*, 587 F. Supp. 2d 114, 115-16 (D.D.C. 2008) (finding that request had been constructively denied after FDA failed to provide him with a determination and stopped replying to his letters).

In subsequent email deliberations between the parties' counsel, Defendants' counsel has since clarified the Census Bureau's extreme and sweeping interpretation of the scope of Title 13's confidentiality provisions: Defendants maintain that beyond withholding personally identifying information, the Bureau must also account for "complementary disclosure" where the release of intermediate data that "does *not appear to contain* individually identifiable information, but *could result* in identifying individuals when those data are coupled with other information in existing Census Bureau publications or other publicly available information." *See* Ex. 3 at 6-9 (7.6.21 Kossak email) (emphasis added). Defendants argue that their all-encompassing approach to confidentiality of any preliminary Census data is necessary because the Bureau "has to keep up with the technology to maintain the public's confidence" in maintaining confidentiality; accordingly, the Bureau "generally avoids the release of intermediate work product [that] can be used in combination with other intermediate work products, official publications, and the final product to re-identify individual respondents and their data items," and thus refuses to release any of Plaintiff's requested intermediate summary or tabulated group quarters data. *See id.*

The problem with Defendants' sweeping interpretation of Title 13's confidentiality provisions, however, is that it conflicts directly with Title 13 itself. As will be shown below, the Bureau's interpretation of Title 13 faces several insurmountable hurdles: the plain language of the disclosure exemptions found in Sections 8 and 9 of the Census Act, together with controlling caselaw and the clear mandates of FOIA itself, all plainly permit the release of *some* Census data, *see* 13 U.S.C. § 8(b), in stark contrast with Defendants' blanket denial of any "intermediate work product" data that "could result" in identifying individuals when combined with other information in existing publications or publicly available information. Besides lacking any factual support beyond Defendants' bald assertion, the Bureau's speculation has no basis in the law itself, and thus cannot support Defendants' sweeping redactions of essentially all "intermediate" group quarters imputation data sought by Plaintiff in the Request. Furthermore, Defendants themselves used an inconsistent method of data redaction in their own 988-page May 25 production, further casting doubt on the supposed grounding of their decisions in the law's mandates. For these reasons, Plaintiff is likely to succeed on the merits of its claim that Defendants misinterpret Title 13's confidentiality provisions, a pure question of law, and thus that Defendants have unlawfully withheld or redacted information subject to disclosure under FOIA.

a. Defendants' withholding of summary and aggregated data in the May 25 production is contrary to Title 13's plain language.

"[T]he starting point for [a court's] analysis is the statutory text." *Desert Palace, Inc. v. Costa*, 539 U.S. 90, 98 (2003). And where, as here, the words of the statute are unambiguous, the "judicial inquiry is complete." *Connecticut Nat'l Bank* v. *Germain,* 503 U.S. 249, 254 (1992) (citation omitted).

The plain language of Sections 8 and 9 of Title 13 unambiguously permits the Secretary of Commerce to release some Census data, including "copies of tabulations and other statistical materials which do not disclose the information reported by, or on behalf of, any particular respondent," 13 U.S.C. § 8(b), as Plaintiff has explicitly requested here, while excluding the underlying raw data originally "*furnished by* any particular establishment or individual" that would identify such an individual, *id.* § 9(a)(2), or "individual reports," *id.* § 9(a)(3), from disclosure or publication. The statute is unmistakably clear in its meaning: it protects the confidentiality of personally identifiable information and raw data as originally furnished by individuals or establishments to the Census Bureau, while permitting disclosure of other tabulations and summary statistical materials that do not disclose such individual information. Defendants' interpretation that these provisions prohibit release of *all* intermediate work product is contradicted by the express statutory provisions.

Although Defendants may argue that Section 9's prohibition on "use [of] the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied," *id.* § 9(a)(1), should be interpreted to protect all intermediate data from release, this interpretation is also atextual and contrary to established canons of statutory construction. While Section 9(a)(1) prohibits the Secretary or other DOC employees' "use" of any furnished information for purposes other than designated statistical purposes, it does not expressly prohibit "publication" of all intermediate information—by contrast, Section 9(a)(2) only prohibits "publication" of personally identifiable data "*furnished* by any *particular establishment or individual*," *id.* § 9(a)(2) (emphasis added), but certainly does not prohibit the "use" of such information internally. Accordingly, when Congress intended to protect certain data from *publication* to third parties, it knew how to do so, and did so explicitly. Principles of statutory interpretation require such provisions instead to be interpreted "to give effect, if possible, to every clause and word," *Duncan v. Walker*, 533 U.S. 167, 174 (2001), and to "avoid statutory interpretations that render provisions superfluous." *United States v. Anderson*, 15 F.3d 278, 283 (2d Cir. 1994). If "use" is construed to mean the same thing as "publication," Section 9(a)(2) becomes entirely duplicative of Section 9(a)(1).

Furthermore, an overbroad interpretation of Section (9)(a)(1) directly conflicts with Section 8(b)'s permitted disclosure of tabulations and other statistical materials that do not contain personally identifiable information. Accordingly, Section 9(a)(1)'s prohibition of the Secretary's "use" of "the information furnished" under Title 13 for non-statistical purposes should be interpreted harmoniously with its surrounding provisions, and not so broadly as to contradict the plain text of the other provisions and to render the rest of the statutory disclosure scheme meaningless. *See Maracich v. Spears*, 570 U.S. 48, 68 (2013) ("The provisions of a text should be interpreted in a way that renders them compatible, not contradictory. . . . [T]here can be no justification for needlessly rendering provisions in conflict if they can be interpreted harmoniously.") (quoting A. Scalia & B. Garner, Reading Law: The Interpretation of Legal Texts 180 (2012)). The judicial inquiry is thus complete on the statutory text alone.

If Congress intended to create an exemption to disclosure for all preliminary Census data, whether tabulations or raw data, it easily could have done so, *see generally Norinsberg v. U.S. Dep't of Agric.*, 162 F.3d 1194, 1200 (D.C. Cir. 1998) ("Had the Congress intended to [incorporate additional statutory requirements], it could have done so expressly."). Indeed, it would have been far simpler to create a blanket confidentiality requirement for all intermediate data (as Defendants assert) than the more detailed and nuanced scheme the statute currently provides. But that is not what Congress did, and this Court should not reinterpret Title 13's express language to match Defendants' preferred understanding of the confidentiality provisions to bar public disclosure about never before used "group quarter imputation," nor should it ignore FOIA's demands to accord with Defendants' preferred disclosure regime.

Importantly, while Section 8(b) uses discretionary, rather than mandatory, language for disclosure—"the Secretary may furnish"—FOIA *requires* that the Bureau promptly furnish any non-exempt responsive records to a FOIA request. Thus, the requested records at issue here, if they exist, must be promptly provided to Plaintiff. 5 U.S.C. § 552(a)(3)(A) ("[E]ach agency, upon any request for records . . . *shall* make the records promptly available to any person." (emphasis added)). Accordingly, any "tabulations and other statistical materials which do not disclose the information reported by, or on behalf of, any particular respondent" must be turned over to Plaintiff in accordance with its FOIA request, even if the data is intermediate work product. In combination, FOIA and Title 13 do not leave room for agency discretion when it comes to withholding such summary statistical materials from a FOIA requester.[15]

b. Defendants' interpretation of Title 13 is inconsistent with controlling caselaw.

As caselaw affirms, Section 8(b) of the Census Act permits the Secretary of Commerce to "furnish copies of tabulations and other statistical materials which do not disclose information reported by, or on behalf of, any particular respondent." *In re England*, 375 F.3d 1169, 1178 (D.C. Cir. 2004) (citation omitted); *see also* 14 Am Jur 2d Census § 9 ("The Secretary of Commerce may also furnish copies of tabulations and other statistical materials which do not disclose the information reported by, or on behalf of, any particular respondent, and may make special statistical compilations and surveys for . . . private persons . . . upon payment of the actual or estimated cost of such work."); *Baldrige v. Shapiro*, 455 U.S. 345, 354-55 (1982) (holding that while "the Secretary [of Commerce] *may* furnish copies of tabulations and other statistical materials which do not disclose the information reported by, or on behalf of, any particular respondent," "raw data *reported by or on behalf of individuals* [is] . . . not available for disclosure" (emphasis added)).

The D.C. Circuit has squarely addressed the question of what data can be disclosed under Title 13, determining that Sections 8(b) and 9(a) permit the Secretary of Commerce to provide

[15] As Plaintiff's counsel communicated to Defendants' counsel on June 29, 2021, many of the redactions that Defendants are insisting upon are demonstrably not protected by Title 13, meaning that FOIA requires they be made promptly available to Plaintiff. For instance, the titles of five fully withheld pages from the production is "County Distribution of 2020 Census – GQ Person Ratios Before and After Imputation." This title demonstrates that the redacted distributions contain imputed group quarters numbers aggregated on a county level, and thus do not disclose confidential raw data reported by particular respondents, making them subject to disclosure under FOIA. *See* Ex. 7 (citing May 25 production). Another redacted page, titled "Summarizing the Map," by its own description includes summary data rather than raw data with personally identifiable information or data reported by individual respondents. *See id.*

"private persons" with "tabulations and statistical materials of a *numerical* nature" in response to FOIA requests, while excluding "names and addresses of specific individuals or firms reporting data to the Census Bureau" for purposes of protecting privacy of individual respondents. *Seymour v. Barabba*, 559 F.2d 806, 809 (D.C. Cir. 1977) (emphasis added). As the Court further explained:

> While a list of names and addresses might be considered to be a "tabulation," yet this would be contrary to the usual understanding. Our understanding of a "tabulation" is a *computation to ascertain the total of a column of figures*, or perhaps counting the names listed in a certain group, rather than supplying the individual names and addresses. This interpretation is made even clearer by the reference in subsection 8(b) to "tabulations and other statistical materials."
>
> We think the authority of the Secretary here to disclose is an authority to disclose numerical statistical data which does not identify any person, corporation, or entity in any way. *Totals*, perhaps *subtotals* and *divisions by categories*, but nevertheless merely *numerical* figures are within this meaning. Individual names and addresses are not.

Id. (emphasis added). Here, because Plaintiff seeks tabulations (*i.e.*, computations of total imputed group quarters data) and statistical materials "of a numerical nature," *id.*, rather than personally identifiable information from the underlying raw data, these records are not exempt from disclosure by the Census Act and must therefore be produced. 5 U.S.C. § 552(a)(3)(A) ("[E]ach agency, upon any request for records . . . *shall* make the records promptly available to any person." (emphasis added)).

In their July 6 email to Plaintiff's counsel, Defendants rely on the Supreme Court's ruling in *Baldrige v. Shapiro* to support their contention that Title 13 exempts essentially limitless swaths of summary-level data from disclosure, thereby exempting such data as is responsive to Plaintiff's FOIA request. Ex. 3 at 7-8 (7.6.21 Kossak email). *Baldrige*, however, bolsters Plaintiff's argument that the data being withheld by Defendants is not raw census data and is therefore ineligible for exemption from disclosure requirements. In *Baldrige*, a county requested disclosure of the Census Bureau's master address register, "a listing of such information as addresses, householders' names, number of housing units, type of census inquiry, and, where applicable, the vacancy status of the unit." 455 U.S. at 349. At issue was whether the register was exempt from disclosure under Exemption 3 and Title 13 sections 8(b) and 9(a), the former of which directs the Secretary to "not disclose the information reported by, or on behalf of, any particular respondent" and the latter of which prohibits publication "whereby the data furnished by any particular establishment or individual under this title can be identified." Such census responses or identifying information are considered raw census data exempt from disclosure. *Id.*

Despite being "compiled initially from commercial mailing address lists and census postal checks," the master address register in *Baldrige* "was updated from data obtained from neighbors and others who spoke with the follow-up census enumerators," meaning it "include[d] data reported by or on behalf of individuals." *Id.* at 358–59. As such, the Court held that the register included raw census data and therefore fell under section 8(b)'s exemption from disclosure.

Similarly, in *Seymour v. Barabba*, the D.C. Circuit reviewed a FOIA request for Census Bureau data including the names and addresses of certain companies. The court held that not only is such information clearly exempt under Title 13, section 9(a)'s prohibition on releasing identifying information, but also that it is separate from the "tabulations and other statistical materials which do not disclose the information reported by, or on behalf of, any particular respondent" that the Secretary may produce under section 8(b). *Seymour*, 559 F.2d at 808-09. In drawing this distinction, the court clarified that Title 13 requires courts to treat individualized identifying information differently from higher-level computations and summaries that do not implicate the same privacy concerns.

Taken together, *Baldrige* and *Seymour* establish that information provided directly by census participants and identifying information such as names and addresses are both exempt from disclosure under Exemption 3 and Title 13 sections 8(b) and 9(a). Plaintiff's Request here, however, asks for neither kind of data. Instead, the Request seeks state-level summaries and tabulations *derived* from the raw census data, representing the kind of higher-level analytical information distinguished by the *Seymour* court as not being exempt under Title 13. Defendants' attempts to cast such non-individualized information as exempt from disclosure is therefore without support in the courts' interpretations of these exemptions. The summary data and aggregate calculations sought implicate none of the same concerns regarding individual participants' identification, as by their very nature they provide only aggregate numbers and general trends. The D.C. Circuit's opinion in *Seymour* particularly indicates such data should be considered separately from identifying information, undermining Defendants' attempt to conflate the two types of data, and further bolstering Plaintiff's case for disclosure under FOIA.

Binding caselaw thus also demonstrates that Plaintiff is likely to succeed on the merits of its FOIA claim—contrary to Defendants' extreme and unfounded position, Plaintiff is entitled to the numerical, summary data and tabulations sought in its Request that is subject to disclosure under Title 13.

c. Defendants' inconsistent approach to redactions further undermines their claims that their particular redactions from the May 25 production are required by Title 13.

Finally, by inconsistently redacting data from the May 25 production, Defendants undermine their own claims that their particular withheld data from the May 25 production were required by Title 13. Instead, their redactions of data appear arbitrary and not governed by standards required by law. To illustrate, in 77 highly similar statistical summary pages for group quarters titled "GQTYPCUR," particular kinds of data (including significant amounts of summary data) are also inconsistently redacted throughout. For instance, some pages have every piece of data fully redacted from the page, while on other surrounding pages Defendants redacted the exact same types of data. *See* Ex. 7. Moreover, on some pages, the range and mean data points were not redacted, where on other pages Defendants partially or fully redacted those data. *See id.*

A consistent approach to redaction should yield uniform applications and outcomes for the same kinds of data; because Defendants' redactions are all over the map on these statistical summary pages, it further undermines Defendants' position that their chosen redactions were required by law. This arbitrariness adds further support for Plaintiff's likelihood of success on the merits.

3. As Defendants implicitly acknowledged by granting Plaintiff's application for expedited processing in the first place, an injunction and order requiring expedited production of improperly withheld portions of its Request is appropriate here to inform an imminent public debate on a matter of national concern.

Plaintiff seeks a preliminary injunction compelling expedited processing of the improperly redacted portions of its FOIA request and that, pursuant to the court's equitable powers to order agencies to act within a particular time frame, *see Landmark Legal Found. v. EPA*, 910 F. Supp. 2d 270, 275 (D.D.C. 2012), this Court order this process to be completed within ten days of the issuing of its order, or before August 15, 2021, whichever is earlier. Defendants argue that a preliminary injunction is not an appropriate vehicle for challenging improperly withheld or redacted records because injunctive relief would be akin to a "decision on the merits." Ex. 3 at 1 (7.16.21 Kossak email). However, this case is akin to the line of precedents where there was an urgency to inform the public of actual or alleged federal government activities, and where courts thus granted preliminary injunctive relief and ordered production of withheld documents.

For instance, as this Court recently recognized in *American Oversight v. U.S. Department of State*, courts in this district have granted preliminary injunctions in cases like this one, and ordered production of non-exempt documents by specific dates, where "FOIA requestors have sought records to inform an imminent public debate on a matter of national concern." 414 F. Supp. 182, 185 n.5 (D.D.C. 2019). *See, e.g.*, *Elec. Privacy Info. Ctr. v. U.S. Dep't of Justice*, 416 F. Supp. 2d 30 (D.D.C. 2006) (ordering release of records related to the Bush Administration's legal justifications for its warrantless wiretapping program in the course of ongoing congressional hearings); *Wash. Post v. U.S. Dep't of Homeland Sec.*, 459 F. Supp. 2d 61 (D.D.C. 2006) (ordering release of records of visitors to the White House and the Vice President's residence *within a period of 10 days* because of the impending midterm elections to be held within a month) (*vacated as moot by subsequent consent motion*, *Wash. Post v. Dep't of Homeland Sec.*, No. 06-5337, 2007 U.S. App. LEXIS 6682, at *1 (D.C. Cir. Feb. 27, 2007)); *Leadership Conf. on Civil Rights v. Gonzales*, 404 F. Supp. 2d 246 (D.D.C. 2005) (ordering release of data regarding the DOJ's responses to election-related civil rights violations in advance of the imminent expiration of the Voting Rights Act).[16]

The Census Bureau's delay in releasing its redistricting population data, arising from the Bureau's novel use of count imputation for tabulating group quarters data, and its publicly acknowledged irregularities resulting from the same, are exactly the kinds of issues subject to current and "imminent public debate on a matter of national concern" justifying an order that Defendants produce non-exempt documents by a date certain.

For instance, the subject matter of Plaintiff's Request has been the subject of widespread media attention. As the Census Bureau has publicly announced, the first release of its "legacy format" summary redistricting data to the states for the 2020 Census has been delayed until August

[16] *See generally Ctr. for Pub. Integrity v. U.S. Dep't of Defense*, 486 F. Supp. 3d 317 (D.D.C. 2020) (ordering release of withheld documents because they were not protected under privileges and therefore not exempt under FOIA).

16, 2021 (with the final release of the redistricting data scheduled to occur on September 30, 2021)[17] to allow for time to address difficulties and irregularities it encountered while gathering and tabulating group quarters data for the 2020 Census, explaining that challenges resulted largely from the COVID-19 pandemic.[18] As described above, in an April 16, 2021 publication on its website by the Chief of the Decennial Statistical Studies Division, the Bureau publicly acknowledged that it "had to adapt and delay some of the ways we counted group quarters because of the COVID-19 pandemic," and explained that, consequently, "[a]fter the end of data collection, when we began processing census data from group quarters, we realized that many of them were occupied on April 1, 2020 (the reference day for the census), but didn't provide a population count."[19] The Bureau also explained the significant impact such group quarters data discrepancies can have for obtaining an accurate population count:

> [W]hen we enumerated [group quarters] in midsummer, some group quarters said they were vacant but they were actually occupied on April 1. If not corrected, such cases could lead to an undercount. If the corrections were not properly coordinated with our procedures to remove duplicated people, they could contribute to an overcount.[20]

Accordingly, the Bureau decided to use some sort of "count imputation" procedures on unresolved group quarters, including for missing characteristics such as age or race, that it "had never before conducted" for group quarters previously.[21]

This data has profound implications for the apportionment of seats in the House of Representatives and the amount of federal funding each State receives for various programs, as well as for the process of drawing electoral districts. It therefore goes without saying that the

[17] Important Dates, U.S. Census Bureau, *supra*.
[18] *Feb. 12, 2021 Census Press Release*, *supra*; *see also Mar. 15, 2020 Census Press Release*, *supra*; *Operational Adjustments*, *supra*.
[19] *Pat Cantwell Statement*, *supra*.
[20] *Id.*
[21] *Id.*

Census Bureau's announcements of irregularities, discrepancies, and novel methods of "count imputation" for group quarters, coupled with the resulting significant delay in releasing the data, is a matter of imminent public debate and national concern. It has also caused many in the media and public to call into question the soundness and reliability of the Bureau's methods, and has thus adversely affected public confidence in the 2020 Census data. For instance, NPR reported that the Census Bureau identified what it described as "processing anomalies" of records for 2020's national tally that "if left unfixed, could miscount millions of people."[22] The report went on to describe how the Bureau had "unearthed major inconsistencies in the information it has gathered this year about residents of college dorms, prisons and other group living quarters—a category that, for the 2020 census, included around 8 million people."[23] Not surprisingly, numerous other large media outlets have covered similar stories of exceptional importance and national interest connected to these irregularities and delays, provoking public skepticism of the accuracy and lawfulness of the Bureau's data collection methods, and questions about whether the resultant delays in redistricting threaten to throw future elections into chaos.[24]

Regarding the urgency of Plaintiff's Request, the time-sensitive nature of it is similar to that in *American Oversight*, where this Court granted the plaintiff's motion for a preliminary

[22] *Millions of Census Records May Be Flawed, supra*; *see also 6-Month Delay in Census Redistricting Data Could Throw Elections Into Chaos*, *supra*.

[23] Wang, *Millions of Census Records May Be Flawed, supra.*

[24] *See, e.g.*, *Mike Schneider, Census Bureau Says Data Irregularities Being Fixed Quickly*, AP (Dec. 3, 2020), https://apnews.com/article/us-news-censuses-census-2020-32fd4322e680365c1de69ab63fc92133 (accessed on July 18, 2021); Reid Wilson, *Census to Delay Data Delivery, Jeopardizing Redistricting Crunch*, The Hill (Feb. 12, 2021), https://thehill.com/homenews/campaign/538649-census-to-delay-data-delivery-jeopardizing-redistricting-crunch (accessed on July 18, 2021); Tara Bahrampour, *Can Americans Trust the Results of the 2020 Census?* Wash. Post (Apr. 26, 2021), https://www.washingtonpost.com/local/social-issues/census-2020-delays-trump/2021/04/22/3a03fbe8-a154-11eb-a7ee-949c574a09ac_story.html (accessed on July 18, 2021).

injunction and ordered release of all expedited, non-exempt documents within in less than one month's time. 414 F. Supp. 3d at 187. Specifically, the plaintiff in that case sought records that went to the heart of an issue Congress was then considering in its inquiry into whether the President of the United States had committed impeachable offenses. *Id.* at 183-84. The Court determined that because the impeachment inquiry was "in full swing" and expected to "conclude by Christmas," time was "clearly of the essence," making the "harm in agency delay . . . more likely to be irreparable." *Id.* at 186-87. Central to this Court's determination there was the fact that, even though the agency had granted expedited processing of the plaintiff's request, the agency had only offered "a very preliminary and incomplete estimate of the number of potentially responsive documents" and had "not even begun to process them." *Id.*

Separately, because Defendants have already granted Plaintiff's application for expedited processing, apparently acknowledging the urgency of its request, *see* Ex. 5, injunctive relief and expedited production of the non-exempt withheld records and data is justified here. To obtain expedited processing under the statute, Plaintiff needed to show a "compelling need," 5 U.S.C. § 552(a)(6)(E)(i), which can be demonstrated by showing, *inter alia*, that the requester is a "person primarily engaged in disseminating information" and that there is an "urgency to inform the public concerning actual or alleged Federal Government activity." *Id.* § 552(a)(6)(E)(v)(II). Alternatively, pursuant to 5 U.S.C. § 552(a)(6)(E)(i)(II), DOC regulations also allow for expedited processing where a request involves "[a] matter of widespread and exceptional media interest involving questions about the Government's integrity which affect public confidence." 15 C.F.R. § 4.6(f)(1)(iii). Defendants' grant of Plaintiff's application for expedited processing is thus an implicit acknowledgment of the accuracy of Plaintiff's argument for a court order of immediate or expedited production of the withheld data. In order is necessary, however, because of Plaintiff's urgency to inform the public about a matter of widespread and exceptional media interest, *i.e.*, the delay and irregularities in the Census group quarters data and its implications for the accuracy of the apportionment count and redistricting.

Because Fair Lines is "primarily engaged in disseminating information" and is likely to succeed in demonstrating an "urgency to inform the public concerning actual or alleged Federal Government activity," 5 U.S.C. § 552(a)(6)(E)(v)(II), this likewise provides an independent basis for granting the preliminary injunctive relief sought by Plaintiff here.

First, while the statute does not define the meaning of a primary disseminator of information, Plaintiff is similar to non-partisan public policy groups that this Court has determined qualify for expedited processing, in that Plaintiff "regularly writes, publishes, and disseminates information." *Brennan Ctr.*, 498 F. Supp. 3d at 9. Indeed, as a Section 501(c)(3) non-profit organization committed to educating the public on fair and legal redistricting, Fair Lines regularly writes, publishes, and disseminates news and information about its comprehensive data gathering, processing, and deployment efforts pertaining to apportionment and redistricting, as well as updating the public on relevant litigation and legal developments throughout the country, while also strategically investing in and publishing academic research. *See* Compl. Ex. C, ECF No. 1-3 at 5-6. In all of these information disseminating activities, Fair Lines is committed to providing public education in the fields of demography, political science, geographic information systems, and legal studies, all while promoting open and transparent government and public accountability by monitoring the activities of policymakers and officials through FOIA requests. *Id.* at 7. Fair Lines uses information gathered from its FOIA requests, and its analysis of that information, to educate the public through reports, press releases, and other media. Fair Lines also publicizes the materials it gathers on its public website and promotes their availability on social media platforms.

Id. Accordingly, Plaintiff is primarily engaged in dissemination of information and meets this statutory requirement.[25]

Second, there is an "urgency to inform the public" here concerning both actual and alleged Federal Government activity, 5 U.S.C. § 552(a)(6)(E)(v)(II), namely concerning the Census Bureau's publicly acknowledged irregularities in the group quarters data for the 2020 Census, especially considering the Census Bureau's unprecedented use of count imputation methods for tabulating unresolved group quarters in the context of the COVID-19 pandemic. This is a matter of current exigency to the American public because these irregularities could lead to gross over- or under-counting of different group quarters populations, and as the media stories and other sources listed above detail, threaten to sow chaos in both redistricting and the upcoming elections. The public has an urgent need to be informed about the reliability (or unreliability) of this data from informed sources like Plaintiff. Furthermore, undue delay threatens to compromise Plaintiff's significant interest in informing the public concerning the soundness and accuracy of the 2020 Census data and imputation processes currently being employed. Finally, given the Census Bureau's impending August 16 release date of the legacy format summary redistricting data to the states for 2020, which will officially commence nationwide redistricting of congressional seats, the public's need to be informed regarding the accuracy of the Census Bureau's data is urgent in every sense of the word.

Accordingly, Plaintiff has demonstrated that this is the kind of exceptional case where injunctive relief regarding redactions and withheld documents from a production is merited; the court should therefore order that Defendants expedite processing of the withheld data under 5

[25] Other courts have found that organizations akin to Fair Lines meet this standard for expedited processing, *see, e.g.*, *Protect Democracy Project v. U.S. Dep't of Def.*, 263 F. Supp. 3d 293, 298-300 (D.D.C. 2017); *Leadership Conf. on Civil Rights*, 404 F. Supp. 2d at 260.

U.S.C. § 552(a)(6)(E)(v).

Furthermore, to obtain preliminary relief by a particular date, Plaintiff must show that it is likely entitled to have the agency finish processing its request by that particular date. *Protect Democracy Project*, 263 F. Supp. 3d at 301; *Brennan Ctr.*, 498 F. Supp. 3d at 99. Because this analysis overlaps significantly with the "irreparable harm" factor, *Brennan Ctr.*, 498 F. Supp. 3d at 99, Plaintiff incorporates here its arguments outlined *infra*.

Defendants' failure to properly expedite processing of Plaintiff's request with respect to the withheld documents is particularly egregious because the agency has publicly acknowledged the unprecedented irregularities that Plaintiff seeks to investigate—the fact that the 2020 group quarters data is known to be highly suspect makes the imminent release of Plaintiff's requested records all the more time sensitive and pressing, because it demonstrates a likely need for action on the part of Plaintiff, the public, and Defendants to correct any errors. Without a court-ordered date for production, Plaintiff and the public "may not otherwise have access" to the records, *Am. Oversight*, 414 F. Supp. 3d, in time for corrective action to be taken by the Bureau before legislative redistricting and the impending elections are fully underway.

Just as the impeachment inquiry at issue in *American Oversight* was in "full swing," warranting expedited production of responsive documents in less than one month's time, *id.* at 186-87, here the Census Bureau has publicly announced that it is currently working to correct problems arising from its group quarters imputation methods and data irregularities. The high stakes of such irregularities in terms of the impact on apportionment and redistricting cannot be overstated. The records sought here pertain to a problem that is not only ongoing, but literally in full swing with the August 16, 2021 legacy data release date drawing near. However, in contrast with the records sought *American Oversight*, Plaintiff's requested records here involve a matter significantly more complicated and intricate: the accuracy of complex statistical methods used to accumulate and tabulate data affecting the count (and by extension, the vote) of potentially millions of individuals living throughout the entire United States. Thus, while a court-ordered production deadline of a little more than a month's time was deemed sufficient in *American Oversight*, Plaintiff here submits that it will need as much time as possible to (1) discover any problems or irregularities with the agency's tabulations or imputation methods, and (2) take action to ensure that corrective measures are implemented by the Bureau in time to prevent irreparable harm to Plaintiff and the public with impending elections that could be directly impacted. As this Court recognized, it is not enough for the public to have awareness of the government's actions of national importance; it is a "structural necessity in a real democracy" that the public have "*timely* awareness" because "stale information is of little value." *Id.* at 186 (quoting *Payne Enters. v. United States*, 837 F.2d 486, 494 (D.C. Cir. 1988)) (cleaned up).

Accordingly, given the time sensitivity of the Census Bureau's impending August 16, 2021 deadline to release the data states need to begin redistricting, Plaintiff requests a court-ordered production deadline of non-exempt responsive data and records from the May 25 production by August 15, 2021, or ten days after the Court's order granting preliminary relief, whichever is earlier. Otherwise, the records requested risk becoming "of little value" to Plaintiff and the public because they will be powerless to do anything to bring about change to the defective process or data, making the harm from inaction more likely to be irreparable, as argued further below. *Id.; see also Brennan Ctr.*, 498 F. Supp. 3d at 101. Defendants' delay in complying with Plaintiff's manifestly compliant Request fully demonstrates the necessity of a preliminary injunction here. The agency has shown no urgency in taking action to fulfill Plaintiff's Request and will likely continue dragging its feet until it will eventually become too late for Plaintiff to take meaningful action and conduct sufficient analysis of the requested records.

Because Plaintiff has shown a likelihood of success on the merits of its claim that Defendants err in their interpretation of Title 13's confidentiality provisions, and thus have withheld non-exempt documents manifestly subject to disclosure under FOIA, this Court should grant Plaintiff's motion for preliminary injunctive relief with an order requiring Defendants to release non-exempt records and data withheld or redacted in response to Plaintiff's Request within 10 days of the Court's order (or no later than August 15, 2021).

B. Plaintiff Will Suffer Irreparable Harm Absent Preliminary Injunctive Relief.

In the absence of a preliminary injunction, Plaintiff will suffer irreparable harm because it will be prevented from receiving information which it is legally entitled to receive under FOIA, and from fulfilling its purpose of informing the public of this matter of highest national concern, until such time as the requested information is no longer relevant or actionable. The D.C. Circuit has established "a high standard" for demonstrating irreparable injury, requiring plaintiffs to show that their injury is "both certain and great; it must be actual and not theoretical." *Chaplaincy of Full Gospel Churches v. England*, 454 F.3d 290, 297 (D.C. Cir. 2006); *Wisc. Gas Co. v. FERC*, 758 F.2d 669, 674 (D.C. Cir. 1985). The core of this inquiry is, of course, the irreparability of the harm. A harm is deemed irreparable when "there can be no do over and no redress." *League of Women Voters v. Newby*, 838 F.3d 1, 9 (D.C. Cir. 2016) (citation omitted). Here, if states are allowed to redistrict using flawed Census Bureau group quarters data, then there would be no conceivable redress for Plaintiff's harm.

Plaintiff has demonstrated that its harm will quickly become irreparable in the absence of judicial intervention. On March 31, 2021, Plaintiff submitted a FOIA request to Defendants seeking records "*deriving from* or summarizing" the responses received to the Census Bureau's

2020 Group Quarters Enumeration questionnaire. Compl., ECF No. 1, ¶ 19. Plaintiff deliberately worded its request in such a way that it is *not* seeking the underlying raw group quarters data that is exempt from disclosure under the Census Act, yet Defendants' first production of records erroneously redacted large swaths of non-exempt information. *See* Exhibit 4. Plaintiff requested, and was granted, expedited processing due to its concern that Defendants would delay providing all documents responsive to the Request for so long that Plaintiff will be unable to adequately inform the public about the content of the requested records and any potential flaws in the group quarters data before the states' redistricting process is in full swing. *Id.* ¶ 20. While the award of expedited processing was welcome, its benefits will not be realized if Defendants continue to illegally withhold the very information needed to accomplish Plaintiff's goal. Plaintiff has raised its objections to specific redactions to no avail, with Defendants standing by their initial determinations despite clear inconsistencies even within the collection of records provided to Plaintiff on May 25. Plaintiff has done all that it can to obtain the release of the improperly withheld information on its own, and without the prompt intervention of this Court Defendants have shown they will not release the group quarters data before Plaintiff suffers irreparable harm from these redactions.

The U.S. Supreme Court has recognized in the context of FOIA that a well-informed public is "a structural necessity in a real democracy," a rule so broad that citizens are generally entitled to the information requested unless an applicable statutory exception applies. *Nat'l Archives & Records Admin. v. Favish*, 541 U.S. 157, 172 (2004). The timing of disclosure also matters a great deal under FOIA. The value of particular data can fluctuate over time, and if a disclosing agency delays release of the requested data for long enough then it can eventually lose *all* of its value. *See Payne Enters.*, 837 F.2d at 494 (noting that "stale information is of little value"). Without a preliminary injunction from this Court, it is likely that the information sought by Plaintiff in its March 31, 2021 Request will eventually lose substantial value because states will commence redistricting without access to that data, new maps will be drawn, and elections will be held based on data that is possibly fundamentally flawed, and no one will be any the wiser. The harmful effects of the states' reliance on potentially faulty group quarters imputation numbers becomes exponentially worse with every passing day, increasing the costs of correcting the numbers and the maps drawn based on them, as states get closer to completing their redistricting processes. Thus, the longer Defendants can stall in turning over any improperly redacted information, the less likely it becomes that any inaccurate apportionment data can be promptly fixed and properly used without massive financial and logistical costs for numerous states throughout the Union. And even if such apportionment errors could eventually be mitigated, additional irreparable harm could result from the public relations havoc that such an unwieldy, delayed exercise could create, leading to reputational damage to our country's electoral system in the eyes of an electorate that is already grappling with doubts about the soundness and integrity of its elections.

The national implications of a defective decennial census make this an exceptional FOIA case, the kind in which "the primary value of the information lies in its ability to inform the public of ongoing proceedings of national importance." *Ctr. for Pub. Integrity v. U.S. Dep't of Def.*, 411 F. Supp. 3d 5, 12 (D.D.C. 2019). Once the Census Bureau publicly releases its legacy format data, with the imputed group quarters numbers baked in, on August 16, the clock of irreparable harm will begin to exponentially speed up as states frantically begin the already delayed redistricting process in preparation for next year's impending election season. Hence, a preliminary injunction is now the only remedy that will adequately protect Plaintiff's right to receive the improperly withheld data at a time when it will still be of full use to Plaintiff.

Plaintiff will suffer irreparable harm in the absence of preliminary injunctive relief because it will be prevented from disseminating important information to the public before a nationwide redistricting process is well underway based on flawed data provided by the Bureau, making the harm more and more difficult to mitigate with each passing day. Plaintiff "seeks records relating to an important public debate and discussion about a process that will come to an end relatively soon." *Brennan Ctr.*, 498 F. Supp. 3d at 103. The entry of a preliminary injunction is therefore necessary to protect Plaintiff (and members of the public that Plaintiff informs) from ongoing, increasing, and irreparable harm.

C. The Balance of the Equities and the Public Interest Favor Granting an Injunction.

The balance of the equities and public interest preliminary injunction factors "merge when the Government is the opposing party." *Nken*, 556 U.S. at 435. Here, any burden that Defendants will incur in terms of processing redacted data and records during an administrative backlog is outweighed by Plaintiff's superior interest in obtaining information and informing the public concerning "an issue of the highest national concern." *Ctr. for Pub. Integrity*, 411 F. Supp. 3d at 15. The injury that Plaintiff will suffer if the requested information is not released in a timely manner is imminent, substantial, and irreparable. Defendants' obstinance with regard to the improperly redacted data indicates that Defendants are uninterested in voluntarily complying with their statutory obligation to produce the requested non-exempt records. In the absence of an injunction that forces Defendants to immediately comply with federal law, Plaintiff will continue to suffer that ongoing and irreparable injury.

Courts weighing the grant of a preliminary injunction "must balance the competing claims of injury and must consider the effect on each party of the granting or withholding of the requested relief." *Amoco Prod. Co. v. Vill. Of Gambell*, 480 U.S. 531, 542 (1987). Unlike Plaintiff, who will suffer a substantial and irreparable harm to their ability to timely inform the public concerning a matter of the greatest public concern, Defendants' only harm from an injunction would come in the form of potential processing delays in other FOIA matters. Although this kind of injury might suffice in some scenarios to tip the balance in Defendants' favor, it does not do so here.[26]

This Court need not look far to find a precedent that is directly on point. The D.C. Circuit recently rejected this very form of injury in a case where a plaintiff sought records related to the 2020 decennial census, the same topic of public concern at issue here. *See Brennan Ctr.*, 498 F. Supp. 3d at 103. In that case, the Court held that the defendant agencies' burden in responding to the FOIA request was "outweighed by the [plaintiff]'s pressing need for the information and the public interest in being informed on a matter—the 2020 census and reapportionment of seats in the House of Representatives—that is *of the highest national concern*." *Id.* (quotation omitted) (emphasis added). Faced with a matter "of the highest national concern," even the increased delays that might result for other FOIA requestors pale in comparison to the harm that Plaintiff will suffer if its request is not timely fulfilled. While "[t]he grant of a preliminary injunction in this case will likely place Plaintiff's request ahead of others in Defendants' FOIA queues, … the extraordinary circumstances presented in this case warrant such line-cutting." *Ctr. for Pub. Integrity*, 411 F. Supp. 3d at 14.

In this case, Defendants' potential harms (such as they are) are outweighed by the overriding public interest in receiving information concerning an issue of vital national importance: the 2020 decennial census and concomitant redistricting of congressional districts

[26] Although Defendants may argue that release of summary data has potential to create harm due to the risk of re-identification of individual respondents and their data items, this "injury" is largely speculative and unsubstantiated. The Court need not take Defendants at their word that such a risk is real or substantial without further elaboration or explanation as to how significant a threat is posed by release of such aggregate, summary-level data that Title 13 does not protect from release.

nationwide. Plaintiff's request is simple; it seeks only information to which it is legally entitled under FOIA. Defendants have demonstrated no inclination to provide the withheld data requested on *any* timetable, much less at a date when it can still usefully inform the public debate over redistricting. In cases where federal agencies refuse to satisfy their obligations under FOIA regarding a matter "of the highest national concern," it is necessary that a court step in to ensure the timely release of the requested information.

Defendants' inexcusable obstruction through these redactions is also contrary to FOIA's express legislative purpose of creating an *expedient* mechanism of providing access to government records. *See Pennsylvania v. United States*, Civil Action No. 05-1285, 2006 U.S. Dist. LEXIS 101810, at *18 (W.D. Pa. Nov. 22, 2006) ("Consistent with the purpose of creating an expedient mechanism for disseminating information and holding government agencies accountable, FOIA directs government agencies to promptly produce any requested materials . . . ."). Even if FOIA's rapid production requirements are demanding for agencies, that is a policy determination Congress made in enacting FOIA's time limits. Indeed, when Congress increased the limit for responding to FOIA requests from 10 days to 20 days, it repeatedly "expressed concerns about agencies delaying their responses" to FOIA requests when doing so. *Beagles v. Watkins*, No. 16-506 KG/CG, 2017 U.S. Dist. LEXIS 143723, at *9 (D.N.M. Sep. 6, 2017). Furthermore, Congress increased the time limits "to make them more realistic," which "signaled the priority Congress placed on agency compliance with the time limits." *Id.*; *Gilmore*, 33 F.Supp. 2d at 1187 (explaining Congress "took these deadlines very seriously" and thus required timely agency responses).

Beyond violating FOIA's textual requirements and contravening its clear purpose, Defendants' obstruction is particularly egregious given the context of the prominence and high stakes of the 2020 Census, concerns the Bureau has publicized regarding its group quarters data collection and imputation methods largely due to the COVID-19 pandemic, and the fast-paced timeframe for apportionment and redistricting, all of which are central to this Request. Practical and judicial economy considerations also favor Plaintiff's requested relief here: absent some form of immediate relief for clear and blatant FOIA violations like these, the message communicated to agencies is that statutory production deadlines and requirements can be circumvented by over-redaction and withholding of documents they do not wish to produce. Agencies will thus have minimal incentive to comply with their FOIA obligations until they are brought into court, further driving up expensive FOIA litigation costs and unnecessarily wasting judicial resources.

Therefore, Plaintiff respectfully requests that this Court grant the requested preliminary injunction to vindicate the Plaintiff's right and the public's interest in the census data requested.

CONCLUSION

For the foregoing reasons, this Court should GRANT Plaintiff's Motion for a Preliminary Injunction and (1) declare that FOIA and Sections 8 and 9 of Title 13 require Defendants to produce tabulations and statistical materials which do not disclose the information reported by or on behalf of any particular respondent, as requested by Plaintiff, including intermediate work product and data without personally identifiable information; (2) order Defendants to immediately identify all non-exempt records and data that were improperly redacted or withheld from the May 25 production;[27] (3) order Defendants to produce all responsive non-exempt records and data from that production within 10 days of the date of the Court's Order, or before August 15, 2021, whichever is earlier; (4) order Defendants to produce all responsive records and data from

[27] To be clear, the Plaintiff is not asserting that every redaction from the May 25 production is improper. Certain redactions—such as internal computer file location descriptions and some of the pages that clearly discuss responses from a single group quarters facility or from individual respondents—are not being challenged in this action.

Plaintiff's proposed narrowed scope of the identified responsive emails from the parties' counsels' negotiations as soon as practicable; and (5) order Defendants to produce a *Vaughn* Index specifically describing in detail each record and portion of each record withheld as exempt within the same timeframe.

Dated: July 19, 2021

Respectfully submitted,

/s/ Jason Torchinsky
Jason Torchinsky (D.C. Bar No. 976033)
jtorchinsky@holtzmanvogel.com
Jonathan P. Lienhard (D.C. Bar No. 474253)
jlienhard@holtzmanvogel.com
Kenneth C. Daines (D.C. Bar No. 1600753)
kdaines@holtzmanvogel.com
HOLTZMAN VOGEL BARAN TORCHINSKY &
JOSEFIAK PLLC
15405 John Marshall Highway
Haymarket, VA 20169
Phone: (540) 341-8808
Fax: (540) 341-8809
Counsel for Plaintiff

CERTIFICATE OF SERVICE

I do hereby certify that, on this 19th day of July 2021, the foregoing Statement of Points and Authorities in Support of Plaintiff's Application for Preliminary Injunction was filed electronically with the Clerk of Court using the CM/ECF system. The system instantaneously generated a Notice of Electronic Filing which served all counsel of record.

/s/ Jason Torchinsky___________
Jason Torchinsky (D.C. Bar No. 976033)
jtorchinsky@holtzmanvogel.com
Jonathan P. Lienhard (D.C. Bar No. 474253)
jlienhard@holtzmanvogel.com
Kenneth C. Daines (D.C. Bar No. 1600753)
kdaines@holtzmanvogel.com
HOLTZMAN VOGEL BARAN TORCHINSKY & JOSEFIAK PLLC
15405 John Marshall Highway
Haymarket, VA 20169
Phone: (540) 341-8808
Fax: (540) 341-8809
Counsel for Plaintiff

EXHIBIT 1

Declaration of Adam Kincaid

UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA

FAIR LINES AMERICA FOUNDATION, INC., Plaintiff, v. UNITED STATES DEPARTMENT OF COMMERCE and UNITED STATES BUREAU OF THE CENSUS, Defendants.	Case No. 1:21-cv-1361-ABJ **DECLARATION OF ADAM KINCAID**

DECLARATION OF ADAM KINCAID

I, Adam Kincaid, declare, pursuant to 28 U.S.C. § 1746 and under penalty of perjury, that the following is true and correct to the best of my knowledge:

1. I am over 18 years of age, a resident of the Commonwealth of Virginia, and competent to testify.

2. I am submitting this declaration in my capacity as Executive Director of Fair Lines America Foundation, Inc. ("Fair Lines").

3. I have personal knowledge of the matters and facts set forth below, the matters and facts set forth in the Statement of Points and Authorities in Support of Plaintiff's Motion for Preliminary Injunction, as well as the matters and facts set forth in the Complaint Fair Lines has brought against the United States Department of Commerce and the United States Census Bureau in the above captioned case filed in the United States District Court for the District of Columbia. Compl., ECF No. 1. To the best of my knowledge, the matters and facts set forth in each of these filings are true and accurate, and the exhibits attached to

these filings are true, complete, and accurate copies of the original documents as represented in these filings.

4. Fair Lines is a Section 501(c)(3) non-profit organization interested in openness and transparency in government, with an emphasis on educating the public and ensuring fair and legal enumeration, apportionment, and redistricting processes. To that end, Fair Lines seeks to review and publicize records in the possession of Defendants in light of the Census Bureau's recent public announcements that it has encountered difficulties and various irregularities regarding the gathering, counting, and imputation of group quarters data for the 2020 Census due to the COVID-19 pandemic and other complicating factors. Fair Lines aims to use these records to fulfill its mission of educating the public about the Census Bureau's activities and their impact on the 2020 apportionment.

5. On February 19, 2021, Fair Lines submitted a FOIA request to the Census Bureau requesting records demonstrating or reflecting the number of residents reported by housing facilities nationwide in response to the Census Bureau's 2020 Group Quarters Enumeration questionnaire. *See* Compl. Exh. A, ECF No. 1-1. On March 12, 2021, Fair Lines received a letter from the Census Bureau denying its request, which asserted that the requested records were exempt from disclosure under 13 U.S.C. § 9 of the Census Act. *See* Compl. Ex. B, ECF No. 1-2.

6. In response to the Census Bureau's denial, on March 31, 2021, Fair Lines submitted a revised FOIA request ("the Request") clarifying that Fair Lines only seeks summaries, tabulations, and other statistical materials derived from, summarizing, or otherwise relating to the original underlying group quarters population data reported for the 2020 Census, rather than the underlying raw data itself. *See* Compl. Ex. C, ECF No. 1-3. In the Request,

Fair Lines clarified that it does not "seek disclosure of the underlying raw group quarters population data itself as originally 'reported by, or on behalf of, any particular respondent' to the Bureau, 13 U.S.C. § 8(b)," nor "any 'publication whereby the data furnished by any particular establishment or individual under this title can be identified,' 13 U.S.C. § 9(a)(2)."

7. In the Request, Fair Lines also included an application for expedited processing of the Request based on its compelling need for the records and the urgency of informing the public of any irregularities in Census Bureau data given the time-sensitive nature of the redistricting process before the impending election season, as well as the decennial nature of the Census Bureau's data collection. Finally, Fair Lines requested a fee waiver or limitation of fees because the records are likely to contribute significantly to public understanding of the operations of the Government and is for non-commercial purposes.

8. On April 7, 2021, having received no confirmation that the Request was received by the Census Bureau, Fair Lines, through its counsel, sent an email to the Census Bureau inquiring about the status of the Request. *See* Compl. Ex. D, ECF No. 1-4. The Census Bureau subsequently affirmed that the Request had been received and that a search had commenced. However, by April 28, 2021, Fair Lines had still received no determination from Defendants regarding the Request, even though the statutory period of twenty business days from the date Fair Lines emailed the Request to the Census Bureau had expired. *See* 5 U.S.C. § 552(a)(6)(A)(i).

9. By May 18, 2021, Fair Lines had still not received a determination from Defendants regarding the Request; accordingly, Fair Lines filed a complaint with this Court on that day. Compl., ECF No. 1. Soon after the Complaint was filed, Fair Lines' counsel received

an email explaining that the Census Bureau was "diligently working" on the FOIA request. Ex. 2. On May 19, the Bureau separately indicated that "in order to conduct an email search for this request, we will need a date range for the emails to search." *Id.*

10. On May 25, 2021, Defendants sent a letter to Fair Lines' counsel (dated May 24, 2021) partially granting and partially denying its FOIA request, and providing Fair Lines with 988 pages of responsive records, Ex. 4; of those, 166 pages were either fully or partially redacted pursuant to FOIA Exemptions 3 and 5, *see id.* Many of these redactions appear to withhold summary or aggregated state- and county-level group quarters data that is of the highest relevance and importance to Fair Lines, and which Fair Lines expressly targeted in its Request. Oddly, Defendants included no records from 2021 in their production. On May 28, 2021, Defendants granted Fair Lines' request for expedited processing of the Request. Ex. 5.

11. Since that time, both parties' counsel have met to discuss these redactions over the phone and email; to the best of my knowledge, the description of communications between both sides' counsel contained in the Statement of Points and Authorities in Support of Plaintiff's Motion for Preliminary Injunction and supporting exhibits are true and accurate. *See* Ex. 3.

12. On July 6, 2021, Defendants' counsel produced two post-December 2020 responsive records that were not included in the May 25 production, but defended all of Defendants' redactions from that production. Ex. 6. To date, Defendants continue to decline to turn over any of the withheld information and data from the 166 redacted pages to Plaintiff.

13. Additionally, Defendants' counsel has indicated that Defendants identified 25,899 pages of emailed material that is potentially responsive to the Request, but have only offered to

attempt review of 300 pages per month for *potential* release to Fair Lines. *See* Ex 3 at 008. On July 10, 2021, Fair Lines' counsel proposed narrowing the scope of the emails to search for records of highest importance to Plaintiff, *i.e.*, documents identifying the total population imputed statewide by the Census Bureau for group quarters. *Id.* at 006. However, on July 16, Defendants' counsel denied this request because the information sought is considered by Defendants to be entirely covered by Title 13's confidentiality provisions. *Id.* at 001. To date, none of these 25,899 pages of emails have been released to Fair Lines, even though the Request for these records was filed on March 31, 2021.

Further Affiant Sayeth Not. Executed on this 19th day of July, 2021.

Adam Kincaid

Exhibit 2

Census Bureau’s Post-Complaint Email
Correspondence to Plaintiff

From: Census EFOIA (CENSUS/PCO) <census.efoia@census.gov>
Date: Tuesday, May 18, 2021 at 2:34 PM
To: Jason Torchinsky <jtorchinsky@hvjt.law>
Subject: Torchinsky_DOC-CEN-2021-001311

Good Afternoon,

We are diligently working on your FOIA request.

Thanks,

Have a great day!

Shauvez Bennett
FOIA Analyst

From: Census EFOIA (CENSUS/PCO) <census.efoia@census.gov>
Sent: Wednesday, May 19, 2021 11:11 AM
To: Jason Torchinsky
Subject: Torchinsky_DOC-CEN-2021-001311

Good Morning,

We are diligently working on your FOIA request. However, in order to conduct an email search for this request, we will need a date range for the emails to search.

Please provide this date range in order to further process this portion of your request.

Very Respectfully,

Shauvez Bennett
FOIA analyst

EXHIBIT 3

Correspondence Between Parties'
Counsel

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Sent: Friday, July 16, 2021 9:59 AM
To: Jason Torchinsky; Ken Daines
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jason and Ken,

Thank you for your proposal to substantially narrow the scope of your request for emails to focus on those most needed by your client. As you stated in your email on July 10, your client requests "only documents identifying the total population (number of individuals) imputed statewide by the Census Bureau for group quarters. We seek these group quarters totals, both resolved and unresolved, tabulated by state. To be clear, we don't request county-level or local-level numbers—only state-level group quarters imputation figures. We also do not seek any household imputation numbers, or numbers reflecting demographic factors like age, race, or sex." You followed up on July 12 to note that the "that information must have been finalized before the state population totals were announced in mid-April, so I believe the timeframe when that document would have been produced internally would be sometime in the 90 days between mid-January and mid-April."

As Defendants understand your request, it seeks information that Defendants have no reason to believe would appear in email. The "group quarters totals, both resolved and unresolved, tabulated by state" that you request is information that Defendants consider to be covered by Title 13's confidentiality provisions. (NB: It would help to understand what your definition of "resolved and unresolved" is so that we can be certain about that). Assuming our understanding of "resolved and unresolved" is consistent with yours, Defendants have no reason to believe that such information was transmitted over email. Rather that information was kept on a secure database. Accordingly, Defendants do not believe it would be fruitful to use search terms to comb through emails that are unlikely to contain the narrowed information your client has requested. Rather, Census could identify the information your client seeks from that database, and would, in all likelihood, withhold it in full pursuant to Exemption 3 for the same reasons I've articulated previously.

If you agree that an email search would be fruitless, we could avoid briefing a preliminary injunction motion and move straight to briefing the merits of Defendants' application of Exemption 3 on summary judgment. While I recognize that you wish to pursue a preliminary injunction, we remain of the view that summary judgment is the appropriate avenue for obtaining a decision on the merits. I would be happy to discuss a summary judgment briefing schedule as an alternative to your contemplated preliminary injunction motion and email search.

As I mentioned previously, I have to fly out to a wedding this afternoon and my email access over the weekend will be spotty. But I am happy to get on the phone on Monday to discuss these issues if you'd like.

- Jonathan

From: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Sent: Tuesday, July 13, 2021 3:13 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>; Ken Daines <KDaines@HoltzmanVogel.com>
Subject: Re: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Thank you. I'll be on a plane much of tomorrow, but Ken will be "on the ground."

- Jason

Jason Torchinsky
Holtzman Vogel Baran Josefiak Torchinsky PLLC

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Date: Tuesday, July 13, 2021 at 1:50 PM
To: Ken Daines <KDaines@HoltzmanVogel.com>
Cc: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Ken and Jason,

Just to close the loop on the below, I was able to confirm that Defendants will not oppose your seeking leave to file a reply on July 30, assuming you file on Monday, July 19, and we oppose on July 26. I'll be in touch regarding your proposal to narrow the email search.

- Jonathan

From: Kossak, Jonathan (CIV)
Sent: Monday, July 12, 2021 8:17 PM
To: 'Ken Daines' <KDaines@HoltzmanVogel.com>
Cc: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Ken,

Sorry for the confusion. When I initially suggested a briefing schedule, I contemplated you filing today or tomorrow, our responding by Friday, July 23, and you selecting a date on which you wanted to reply. Per your email below, however, if you will file on Monday, July 19, my opposition will be due on Monday, July 26, so there is no need for a briefing schedule. I'll double-check with Census to make certain that they would not oppose your seeking leave to file a reply on July 30 and should be able to get back to you tomorrow.

- Jonathan

From: Ken Daines <KDaines@HoltzmanVogel.com>
Sent: Monday, July 12, 2021 5:31 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Cc: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jonathan,

We would be amenable to filing on Monday, along with a briefing schedule as you suggest. Could you please send us a draft proposed briefing schedule, with the dates as follows:

- Plaintiff's PI motion due Monday, July 19
- Defendants' opposition due Monday, July 26
- Plaintiff's reply due Friday, July 30

Thank you,

Ken

Ken Daines
KDaines@HoltzmanVogel.com // www.HoltzmanVogel.com

Holtzman Vogel
HOLTZMAN VOGEL BARAN TORCHINSKY & JOSEFIAK PLLC

PRIVILEGED AND CONFIDENTIAL
This communication and any accompanying documents are confidential and privileged. They are intended for the sole use of the addressee. If you receive this transmission in error, you are advised that any disclosure, copying, distribution, or the taking of any action in reliance upon this communication is strictly prohibited. Moreover, any such disclosure shall not compromise or waive the attorney-client, accountant-client, or other privileges as to this communication or otherwise. If you have received this communication in error, please contact me at the above email address. Thank you.

DISCLAIMER
Any accounting, business or tax advice contained in this communication, including attachments and enclosures, is not intended as a thorough, in-depth analysis of specific issues, nor a substitute for a formal opinion, nor is it sufficient to avoid tax-related penalties. If desired, Holtzman Vogel Josefiak Torchinsky, PLLC would be pleased to perform the requisite research and provide you with a detailed written analysis. Such an engagement may be the subject of a separate engagement letter that would define the scope and limits of the desired consultation services.

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Sent: Monday, July 12, 2021 5:08 PM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Cc: Ken Daines <KDaines@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jason,

Thank you very much for your consideration of my schedule. Will you be sending me your proposed filing today (even if you don't file until Friday)? Otherwise, if you file on Friday, I will still lose two days of my response time over the weekend. If you don't plan to share it beforehand, would you be amenable to filing on Monday? That way, my opposition will be due on Monday, July 26, and I'll have the full seven days to oppose. Also, are you going to seek leave to reply? If so, it seems like proposing a reasonable briefing schedule is still the best way to go.

- Jonathan

From: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Sent: Monday, July 12, 2021 4:53 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Cc: Ken Daines <KDaines@HoltzmanVogel.com>
Subject: Re: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jonathan,

Thanks for the response.

On the PI motion, in light of your schedule, I think we'll plan to file this Friday, and that would make your response due July 23rd without needing to seek permission from the court.

Please let us know when you are able to provide some feedback from your client on the remaining document search. Just FYI – that information must have been finalized before the state population totals were announced in mid-April, so I believe the timeframe when that document would have been produced internally would be sometime in the 90 days between mid-January and mid-April.

Thanks,
Jason

Jason Torchinsky
Holtzman Vogel Baran Josefiak Torchinsky PLLC

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Date: Monday, July 12, 2021 at 4:13 PM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Cc: Ken Daines <KDaines@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jason,

I am not in a position to discuss your separate proposal regarding the narrowing of the emails yet. On the PI Motion, you are correct that we oppose your filing. When are you planning on filing? I ask because under the Local Rules, we would only have 7 days to file our opposition. I am leaving for a family wedding this Friday (yes, another one – my only two of the summer), and I would appreciate having until Friday, July 23 to respond. In exchange for this courtesy, I am sure that Census would not oppose your filing a reply (one is not permitted as of right) on a reasonable time-table.

Please let me know if you are willing to agree to a briefing schedule of this sort. I'm happy to discuss.

- Jonathan

From: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Sent: Monday, July 12, 2021 2:44 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Cc: Ken Daines <KDaines@HoltzmanVogel.com>
Subject: Re: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jonathan,

Just to be clear, pursuant to DDC Local Rule 7(m), it appears that Defendants oppose Plaintiff's filing of a PI motion. Is that correct? We clearly have a fundamental disagreement over the relief that the Court in its discretion is able to afford via preliminary injunction, and of the scope of protected materials under Title 13.

If you would like to discuss our separate proposal described in my email below to narrow down the 917 emails/attachments, we can schedule a phone call for that.

-Jason

Jason Torchinsky
Holtzman Vogel Baran Josefiak Torchinsky PLLC

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Date: Monday, July 12, 2021 at 8:01 AM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Cc: Ken Daines <KDaines@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jason,

I was at a wedding this weekend and did not anticipate your Saturday afternoon message. I'd like to discuss your email with my client this morning (if I'm able) and have a call with you in the afternoon to discuss.   After that, you will obviously be free to do as you wish. But just for clarification, in our last meet-and-confer, you appeared to move away from filing a preliminary injunction and towards a partial motion for summary judgment. Have you gone back to thinking about a preliminary injunction?   You seemed to want a determination on the merits, which is not what a PI will get you.

- Jonathan

From: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Sent: Saturday, July 10, 2021 4:06 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Cc: Ken Daines <KDaines@HoltzmanVogel.com>
Subject: Re: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jonathan,

We appreciate your client's position, but please understand that our client has significant and real concerns about the Constitutional command to conduct an actual enumeration. Your client appears to have used a method of "imputation" never before applied to Group Quarters to establish numbers used to determine the apportionment of Congressional seats between the states. This is both significant and time sensitive, particularly in light of the Census Bureau's impending August release of the legacy format summary data and the redistricting process that will commence in earnest immediately afterward. While your client seems to believe it can make these significant determinations without sunlight and judicial oversight, I believe your client is mistaken.

In addition, while your client explained its recent re-interpretation of Title 13's privacy requirements to the court in Alabama, that court did not yet address the merits of your client's position. The denial of the preliminary injunction did not address the substance of either side's positions.

We assume from your email below you will oppose our forthcoming motion for preliminary injunction, which will challenge your client's improper redactions from the current production. We look forward to briefing and argument about the proper scope of Title 13's confidentiality provisions and the urgency of this request.

Separately, we appreciate your offer to negotiate the parameters for searching the remaining 917 potentially responsive emails and attachments—we propose substantially narrowing the scope of the universe of emails to focus on those most needed by our client. Specifically, our client requests narrowing the email search to only seek documents identifying the total population (number of individuals) imputed statewide by the Census Bureau for group quarters. We seek these group quarters totals, both resolved and unresolved, tabulated by state. To be clear, we don't request county-level or local-level numbers—only state-level group quarters imputation figures. We also do not seek any household imputation numbers, or numbers reflecting demographic factors like age, race, or sex. Please let us know if your client agrees to this proposal, and what the estimated production timeline would be.

Thank you,

Jason

Jason Torchinsky

Sent from my mobile device. Please excuse any typos.

> On Jul 6, 2021, at 9:32 PM, Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov> wrote:
>
> Dear Jason and Ken,
>
> Thanks for your email last week. As you note, Defendants redacted only 115 of the 988 pages they produced on May 24, 2021, in response to your client's FOIA request in this litigation. As I explained in my June 25, 2021 email, the redactions

were made by the Census Bureau’s Disclosure Review Board (DRB), whose purpose is to support the Data Stewardship Executive Policy (DSEP) Committee to ensure that every information product released by the Census Bureau adheres to the confidentiality requirements of Title 13 and other applicable statutes. As you are aware from the face of your client’s request, 13 U.S.C. §§ 8(b) and 9 are the statutory provisions under the Census Act that impose a mandate upon the Census Bureau to protect the confidentiality of individual census responses and data. These provisions prohibit the Census Bureau from releasing “any publication whereby the data furnished by any particular establishment or individual under this title can be identified,” and allows the Secretary to provide aggregate statistics so long as those data “do not disclose the information reported by, or on behalf of, any particular respondent.”

Other than the inconsistency you purport to identify in the first bullet of your email, the remainder of your concerns appear to be driven by your misconception of how the Title 13 confidentiality provisions work. You contend that the DRB improperly redacted certain data because “it is only derived from raw data, but does not include the numbers that were furnished by any particular establishment or individual”; or that certain “statistical information or tabulations . . . do not disclose any raw data reported by particular respondents”; or that certain “categories of data described are clearly summary in nature, and would not lead to disclosure of any particular respondent’s reported data.” These arguments, and those repeated in the same or similar wording in your other bullets, are all based upon the same erroneous conception of Title 13’s confidentiality provisions.

As you are aware from the State of Alabama litigation in which you participated, to satisfy Title 13’s privacy strictures, the Census Bureau must account for “complementary disclosure,” which is the release of data that does not appear to contain individually identifiable information, but could result in identifying individuals when those data are coupled with other information in existing Census Bureau publications or other publicly available information. As you are also aware from the Alabama case, the Census Bureau has dedicated significant resources to addressing the Fundamental Law of Information Reconstruction, which says that overly accurate estimates of too many statistics can destroy privacy. Modern computational and information resources feed on statistical data, and the cumulative effect of statistical releases in this age of computing power and sophistication poses a significant threat to the privacy of individual responses. The Census Bureau generally avoids the release of intermediate work product because it can be used in combination with other intermediate work products, official publications, and the final product to re-identify individual respondents and their data items.

The DRB reviewed the 988 pages produced to you and determined that the withheld data had to be redacted because its release would violate Title 13's confidentiality provisions in light of complementary disclosure and/or reconstruction concerns. I know of no FOIA case (nor any other case in any other context) that undermines the Census Bureau's authority to redact this information. Indeed, the last significant challenge in the context of FOIA to the Census Bureau's withholding of information pursuant to Title 13's confidentiality provisions was *Baldridge v. Shapiro*, 455 U.S. 345 (1982), in which the Supreme Court reviewed the history of those provisions and determined that Congress's intention in establishing the confidentiality provisions, "was to encourage public participation and maintain public confidence that information given to the Census Bureau would not be disclosed." *Id*. at 361. *Baldridge* is nearly 40 years old and technology has greatly advanced since then. The Census Bureau has to keep up with the technology to maintain the public's confidence. Title 13's confidentiality provisions would be severely undermined if the Census Bureau did not take into account the risk of re-identification attacks on aggregated data releases. Accordingly, the redactions you identify below are not "improper." We are confident they will stand against challenge in any court.

However, as I mentioned on our last call, any such challenge is premature. Motions for partial summary judgment in FOIA cases are heavily disfavored by the courts in this jurisdiction, and you have not identified any particular reason why the redacted data is needed urgently. Moreover, you already have received the vast majority of information in an unredacted manner, and the Census Bureau will be publicly releasing vast quantities of data no later than August 16, 2021. Your client has asked for emails responsive to its FOIA request, and Defendants have identified 917 potentially responsive emails, consisting of 25,899 pages of material. That does not include either attachments to those emails or Excel spreadsheets. The attachments increase the number of documents to 2,414 and the page count to 35,880 pages. The Excel spreadsheets, which would be produced in native format, have to be converted into pdfs to get a page count. The total page count figure for the excel spreadsheets would be 760,000. That is obviously an astronomical figure. In the ordinary course of a FOIA litigation, we would work with a plaintiff to figure out how to narrow the universe of potentially responsive material down to reasonable proportions, but that takes time. As stated, Defendants will use their best efforts to process 300 pages of potentially responsive records every month. It may be that in 2-4 months your client determines that "the juice is not worth the squeeze," and agrees to forego further processing. Or your client may identify certain materials in the disclosed records that it finds useful and may agree to narrow the universe of material to be reviewed. We are happy to continue negotiating the parameters of your request, but such negotiations are likely to be more productive after a few months of processing have taken place.

Given the early stage of this litigation, we intend to oppose as premature any motion for partial summary judgment you seek leave to file. And even if the Court allows it, we will move to stay the processing of any additional records until after the briefing process is complete, since that process will take up the resources of key staff who would otherwise be participating in the processing of potentially responsive records.

Finally, attached are the two additional "post-December 2020" documents we have been discussing in the emails below and in our last call. As for your concern that it seems unlikely that there are only two such documents, the Census Bureau has verified for us that the documents produced are the only ones responsive to your FOIA request. For your awareness, Defendants have employed the typical "date-of-search" temporal limitation blessed by the D.C. Circuit. For the post-December 2020 records, the date the search for those records began was May 19, 2021.

I'm happy to discuss any of the above in more depth this week. Please let me know when you are available.

- Jonathan

From: Ken Daines <KDaines@HoltzmanVogel.com>
Sent: Tuesday, June 29, 2021 8:06 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Cc: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jonathan,

As we discussed, I am attaching a pdf with 115 redacted pages pulled from the Bureau's 991-page production where it is most apparent (and in several cases indisputable) that summary statistical information was improperly redacted. Without providing an exhaustive description of our rationale for challenging each page, here are some examples where redaction under Title 13 was improper (along with corresponding page numbers from the pdf we are providing):

- **GQTYPCUR Statistical Summary Pages (pp. 1-77):** Here it is clear that statistical summary data is redacted, including the Min, Q1-3, Max, and in some cases the Mean, Range, and Std Dev. What appears to be histograms at the bottom of each page are also improperly redacted. The information from these pages are improperly redacted under 13 U.S.C. § 8(b) because it is only *derived from* raw data, but does not include the numbers that were furnished by any particular establishment or individual to the Bureau, and would not lead to disclosure of such data or include identifying information. Furthermore, the data is inconsistently redacted, suggesting that an arbitrary method was used; for instance, on page 44, every piece of data is redacted, even though the same types of data on the previous and subsequent pages are not redacted. On some pages the range and the mean are fully included, while other pages have them partially or fully redacted.

- **County Distribution of 2020 Census – GQ Person Ratios Before and After Imputation (pp. 79-82, 104-105) –** The title of these pages makes clear that group quarters distribution numbers are shown on the county level, including summary statistical information or tabulations that do not disclose any raw data reported by particular respondents.
- **Pages 83-89 –** Redacted information includes summary statistical information that is not the originally reported raw data, including Mean, Std. Dev, Minimum, Maximum, and Median, Mode, 25th and 75th Percentiles.
- **Pages 90-91** – Histograms are redacted, but no reason to believe these include raw data reported by particular respondents.
- **Group Quarters Imputation Methodology (p. 92)** – "Median Good People Count" is summary or tabulated data, not data that was originally reported or identifying data.
- **District of Columbia and South Carolina tables/charts (pp. 94-95)** – The categories of data described are clearly summary in nature, and would not lead to disclosure of any particular respondent's reported data. E.g., for D.C. it includes a row titled "2020 DRF1 Total Population" that is improperly redacted.
- **"Summarizing the Map" (p. 97)** – The numbers in this document by its own description, "summarizing," are nothing more than summary data. E.g., one redacted number pertains to the number of tracts that have a percentage decline of 90% or more, etc. But none of these include raw data as it was reported by individual respondents.
- **Census Tracts with 100% Decline from 2013-2017 ACS (p. 98)** – Here the Bureau could provide the state-, county-, and tract-level information while omitting the identifying facility names. The same is true for other pages with Census tracts data, including **pages 100-101**.
- **Pages 106-108** – These also appear to be summary statistics based on the table format, although it is admittedly difficult to tell based on the full redaction.
- **Tracts with Largest Number of Nursing Home People Found in a GQ (pp. 109-114)** – The state-, county-, and tract-level data is summary statistical information that does not disclose information reported by any particular respondent.
- **10 Counties with Highest % Enrolled (p. 115)** – The Bureau can provide the percentage, county- and state-level information, without providing particular university information.

Please note that by providing these examples, including the pdf, we are not waiving our right to challenge improper redactions on the other redacted pages, many of which are fully redacted which makes it impossible to tell whether redaction was improper.

Also, as discussed on the call, we look forward to your update this week regarding the post-December 2020 documents and the 2600 emails (including the number of pages).

Thank you,

Ken

Ken Daines
KDaines@HoltzmanVogel.com // www.HoltzmanVogel.com

--

PRIVILEGED AND CONFIDENTIAL
This communication and any accompanying documents are confidential and privileged. They are intended for the sole use of the addressee. If you receive this transmission in error, you are advised that any disclosure, copying, distribution, or the taking of any action in reliance upon this communication is strictly prohibited. Moreover, any such disclosure shall not compromise or waive the attorney-client, accountant-client, or other privileges as to this communication or otherwise. If you have received this communication in error, please contact me at the above email address. Thank you.

DISCLAIMER
Any accounting, business or tax advice contained in this communication, including attachments and enclosures, is not intended as a thorough, in-depth analysis of specific issues, nor a substitute for a formal opinion, nor is it sufficient to avoid tax-related penalties. If desired, Holtzman Vogel Josefiak Torchinsky, PLLC would be pleased to perform the requisite research and provide you with a detailed written analysis. Such an engagement may be the subject of a separate engagement letter that would define the scope and limits of the desired consultation services.

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Sent: Friday, June 25, 2021 9:11 PM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Cc: Ken Daines <KDaines@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Yes, I can make that time. I think for our last call you invited me to a Zoom meeting. I'm happy to do that again, or you can call me at 202-598-5772.

- Jonathan

From: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Sent: Friday, June 25, 2021 8:36 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Cc: Ken Daines <KDaines@HoltzmanVogel.com>
Subject: Re: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jonathan,

Would 11:30am Tuesday morning ET work for you for a call? There are several things below that I think are resolvable with some discussion.

Thanks,
Jason

Jason Torchinsky
Holtzman Vogel Baran Josefiak Torchinsky PLLC

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Date: Friday, June 25, 2021 at 4:01 PM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Cc: Ken Daines <KDaines@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jason,

Thanks for your email. I disagree with parts of your characterization of our communications, but putting that to the side, Defendants would oppose your motion for preliminary injunction (PI) and I really don't see any basis you have to file such a motion. You previously threatened a PI at the end of May, but relented when you received Defendants first production of nearly 1,000 pages. You still have that production, a vast majority of which (over 80%) is unredacted. I'm not sure what has changed to provide you with a basis for seeking a PI. What you have asked is for Defendants to identify for each redaction specifically which FOIA Exemption justifies the redaction and also to reconsider the redactions. Defendants have considered your request and determined the following:

All information withheld was redacted by the Disclosure Review Board (DRB), whose purpose is to support the Data Stewardship Executive Policy Committee to ensure that every information product released by the Census Bureau adheres to the confidentiality requirement of Title 13 and other applicable statutes. Specifically, the data redacted on pages 339-415, 428, 430-437, 443, 450, 457, 467-468, 474-475, 524, 533, 539, 555, 574, 596, 607, 628, 639, 731-732, 746-753, 757, 877, 916-918, 927-928, 930-934, 939-945, 949-950, 955-963, 972-974, 976-978, and 988 was all determined by the DRB to constitute Title 13 information that cannot be disclosed and is therefore covered by FOIA Exemption 3. Defendants stand by all of these redactions. Additional FOIA Exemptions may apply, but Defendants are still in the process of making that determination.

In addition, 18 pages (221, 229, 237, 243, 249, 257, 265, 273, 282, 295, 308, 880-881, 884, 887, and 889-891) contained partial redactions of file names, including internal pathways identifying where secure file information is located. Again, the DRB made the determination to withhold these file names and path structures. Again, Defendants stand by all of the redactions, but are still in the process of determining all of the specific exemptions that apply to those redactions.

Whether these records were properly redacted is a matter for summary judgment briefing, not a preliminary injunction motion. The latter is intended to maintain the status quo, not give your client the relief it seeks on the merits.

As to the email records you requested, as I mentioned, Defendants have completed their search for potentially responsive email records and identified approximately 2600 potentially responsive emails. However, Defendants have not completed the process of threading/deduplication, which is likely to reduce that figure. In the next JSR due on July 20, 2021, Defendants will agree to using their best efforts to process 300 pages of potentially responsive records per month, with the first release of any nonexempt, responsive records by July 30, 2021, and continuing on thereafter on a monthly basis until all the potentially responsive emails are processed. (NB: Processing 300 pages per month does not guarantee production as the records may be found to be exempt, unresponsive, or may require consultation). This is a very standard and reasonable FOIA schedule. It takes time to review records for responsiveness and determine whether any FOIA exemptions apply. And, as you know, Defendants must balance responding to your client's request with their competing responsibilities to other FOIA requesters. I'm happy to discuss this schedule as we put together a joint status report.

Finally, regarding your request that Census provide material similar to that which was provided to you in Defendants' initial production for the post-December 2020 time period, Census has not yet completed its search.

I propose that we speak on Monday or Tuesday (I'm generally available in the mornings on both days), if you still have concerns after digesting this information. I believe such a discussion is more likely to address your concerns than a PI briefing process. Please let me know if you have times available.

Regards,

Jonathan

From: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Sent: Friday, June 25, 2021 6:16 AM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Cc: Ken Daines <KDaines@HoltzmanVogel.com>
Subject: Re: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jonathan,

Given the potentially time sensitive nature of the information contained in these documents and impact on the impending redistricting process, I will be consulting with my client about whether we are going to seek a preliminary injunction. There were only a relatively small number of pages with redactions in the initial production, so I fail to see why this has taken this amount of time. We had this initial discussion awhile ago, a date was promised, and now your client has both unilaterally pushed back that agreement and further slowed the production process.

With respect to the additional documents that we requested and are entitled to, if an initial search is already complete, I fail to see why documents cannot be produced on a rolling basis well in advance of July 20.

I look forward to a prompt response from your client, and your position on whether you would oppose a preliminary injunction motion.

Sincerely,
Jason Torchinsky

Jason Torchinsky
Holtzman Vogel Baran Josefiak Torchinsky PLLC

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Date: Thursday, June 24, 2021 at 11:36 PM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Cc: Ken Daines <KDaines@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jason,

Apologies for the late email. I thought I had sent it earlier, but I went to close down my computer and it was still in my draft folder. I'm going to be out of pocket in the morning, so I'm sending this now. Please excuse the late-night intrusion.

Unfortunately, I don't have any concrete information to offer. Defendants ran into some unexpected technical difficulties and are working through them. They are hoping to have the documents available by the end of the month, but cannot guarantee that at the moment. I will stay on top of this and let you know when I have any updates.

As for the emails, Defendants have done an initial search and are in the process of determining a total page count. Our first JSR is due on July 20, 2021, and I am hopeful that we will know the total page count and be able to negotiate a processing schedule far in advance of that date.

I'll be in touch with any additional updates on both subjects.

- Jonathan

From: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Sent: Thursday, June 24, 2021 2:04 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Cc: Ken Daines <KDaines@HoltzmanVogel.com>
Subject: Re: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jonathan,

Just checking in. Any idea when the documents might be available? And have you been able to get any answers on the additional production?

Thanks,
Jason

Sent from my iPad

On Jun 17, 2021, at 6:43 PM, Kossak, Jonathan (CIV)
<Jonathan.Kossak@usdoj.gov> wrote:

Jason and Ken,

I'm still tracking down an answer on your question about emails, but I wanted to touch base regarding Defendants' due date for answering the complaint. Under FOIA and the Federal Rules of Civil Procedure, Defendants must file an answer within 30 days after service on the U.S. Attorney. The U.S. Attorney's Office for the District of Columbia does not have a record of being served in this case. I recognize that the Summons as filed on the docket says that the Summons for the U.S. Attorney's Office was issued on May 20, 2021, but do you have proof of the date of service? Assuming that you do, 30 days from May 20 is Saturday, June 19, 2021, which means that Defendants' deadline is Monday, June 20, 2021. Accordingly, Defendants plan to file their Answer Monday (assuming you can demonstrate proof of service). Please let me know if this is not consistent with your understanding of Defendants' deadline.

Thanks,

Jonathan

From: Kossak, Jonathan (CIV)
Sent: Wednesday, June 16, 2021 1:10 PM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>; Ken Daines
<KDaines@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

I'll check in on that and will get back to you.

- Jonathan

From: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Sent: Wednesday, June 16, 2021 12:07 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>; Ken Daines
<KDaines@HoltzmanVogel.com>
Subject: Re: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jonathan,

Thanks. How about the additional materials we discussed that have not been produced? Emails and documents after December 2020.

Jason

Jason Torchinsky
Holtzman Vogel Josefiak Torchinsky PLLC

Sent from my phone. Please excuse any typos.

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Sent: Wednesday, June 16, 2021 11:44:17 AM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>; Ken Daines <KDaines@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jason,

Regarding the below discussion, Defendants plan to reprocess and release the records previously released to you by next Thursday, June 24. Of course, if they complete this task prior to that date, I will let you know and send you the material as soon as I have it ready.

Best,

Jonathan

From: Kossak, Jonathan (CIV)
Sent: Thursday, June 10, 2021 2:05 PM
To: 'Jason Torchinsky' <jtorchinsky@HoltzmanVogel.com>; 'Ken Daines' <KDaines@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Hi Jason,

Just to follow up on the below, my client contact had a family medical emergency on Tuesday evening that he is still dealing with. He has promised to get back to me soon with an update, but out of courtesy and respect to his emergency, I would like to give him some space. If I haven't heard back by COB on Monday, I'll follow up on Tuesday and see if I can get some solid information.

Thanks for your patience,

Jonathan

From: Kossak, Jonathan (CIV)
Sent: Tuesday, June 8, 2021 4:39 PM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>; Ken Daines
<KDaines@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Hi Jason,

I have an email in to my client about this. I'm not certain that I will hear back today, but I expect a response by tomorrow. I'll let you know when I can confirm their timing.

- Jonathan

From: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Sent: Tuesday, June 8, 2021 4:11 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>; Ken Daines
<KDaines@HoltzmanVogel.com>
Subject: Re: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jonathan,

Just checking in. Any update on when your client might have more for us?

Thanks,
Jason

Jason Torchinsky
Holtzman Vogel Baran Josefiak Torchinsky PLLC

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Date: Friday, May 28, 2021 at 11:41 AM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>, Ken Daines
<KDaines@HoltzmanVogel.com>
Subject: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jason and Ken,

Please see the attached correspondence from Commerce/Census regarding your client's request in the above-captioned case.

Have a good weekend,

Jonathan

From: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Sent: Thursday, May 27, 2021 2:12 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>; Ken Daines <KDaines@HoltzmanVogel.com>
Subject: Re: LCvR 7(m) Confer - Preliminary Injunction Motion

Jonathan,

Thank you for the prompt response. Please keep us posted.

We'd also like to receive the additional documents on a rolling basis as they are ready for release.

Thanks
Jason

Jason Torchinsky
Holtzman Vogel Josefiak Torchinsky PLLC

Sent from my phone. Please excuse any typos.

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Sent: Thursday, May 27, 2021 1:55:59 PM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>; Ken Daines <KDaines@HoltzmanVogel.com>
Subject: RE: LCvR 7(m) Confer - Preliminary Injunction Motion

Jason,

I just wanted to circle back on this. My clients have agreed to reprocess the production released to you earlier this week to mark the redactions with the applicable exemptions so that it is clear which exemptions they have applied to a particular redaction. They have also agreed to review the redactions to determine whether they stand by those redactions. If my clients determine that they inappropriately withheld information, they will release it upon reprocessing. I don't yet have a timetable on how long that process will take, but I'll let you know as soon as I do.

- Jonathan

From: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Sent: Wednesday, May 26, 2021 9:41 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>; Ken Daines <KDaines@HoltzmanVogel.com>
Subject: Re: LCvR 7(m) Confer - Preliminary Injunction Motion

Mr. Kossak,

Thank you for meeting with us this morning. We look forward to hearing your update on the redactions we discussed, and to working with you to arrive at a resolution.

As we also discussed on the call, in response to the Census Bureau's May 19, 2021 question about the scope of the email search, we are willing to narrow the scope of the search to all responsive emails sent or received **between March 31, 2020 and March 31, 2021**. If any email attachment contains responsive information, summaries, tabulations, etc., please include the original email as well as all attachments to it. Please note that by agreeing to this modification (and by any other statement in this email), we do not waive our right to pursue any remedies requested in our Complaint or otherwise, nor do we waive, toll, or reset the FOIA statutory requirements and deadlines governing this Request.

Regarding your request for a description of the information we are targeting, we are looking for the following in particular:

> Summaries, tabulations, and other statistical materials that demonstrate the aggregate number of individuals (or percentage of the total) that were counted or imputed as part of any 2020 Census enumeration tabulations (whether preliminary or final) as a result of group quarters imputation procedures (i.e., for unresolved group quarters), with numbers aggregated on a statewide level and on a county-wide level for each state. We also seek email or other correspondence that summarizes or identifies the same information, or includes it as an attachment.

Again, as stated in Fair Lines' Request, by requesting these numbers of individuals counted or imputed for unresolved group quarters via imputation by the Bureau, we are not requesting the underlying raw group quarters population data as originally "reported by, or on behalf of, any particular respondent" to the Bureau, 13 U.S.C. § 8(b), nor do we seek any "publication whereby the data furnished by any particular establishment or individual under this title can be identified," 13 U.S.C. § 9(a)(2), or other "individual reports," 13 U.S.C. § 9(a)(3), but rather only *aggregated* numbers on a statewide or county-wide level.

Please let us know if we can answer any questions or provide additional clarity.

Thank you,

Jason

Jason Torchinsky
Holtzman Vogel Baran Josefiak Torchinsky PLLC

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Date: Tuesday, May 25, 2021 at 3:18 PM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>, Ken Daines <KDaines@hvjt.law>
Subject: RE: LCvR 7(m) Confer - Preliminary Injunction Motion

Yes, that works for me. Talk to you then.

- Jonathan

From: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Sent: Tuesday, May 25, 2021 3:12 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>; Ken Daines <KDaines@hvjt.law>
Cc: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Subject: Re: LCvR 7(m) Confer - Preliminary Injunction Motion

Jonathan,

Nice to meet you by email.

Can you talk at 1030am tomorrow? If so, I can circulate a dial in.

- Jason

Jason Torchinsky
holtzmanvogel.com

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

PRIVILEGED AND CONFIDENTIAL
This communication and any accompanying documents are confidential and privileged. They are intended for the sole use of the addressee. If you receive this transmission in error, you are advised that
copying, distribution, or the taking of any action in reliance upon this communication is strictly prohibited. Moreover, any such disclosure shall not compromise or waive the attorney-client, accountant-c
privileges as to this communication or otherwise. If you have received this communication in error, please contact me at the above email address. Thank you.

DISCLAIMER
Any accounting, business or tax advice contained in this communication, including attachments and enclosures, is not intended as a thorough, in-depth analysis of specific issues, nor a substitute for a formal
it sufficient to avoid tax-related penalties. If desired, Holtzman Vogel Josefiak Torchinsky, PLLC would be pleased to perform the requisite research and provide you with a detailed written analysis. Such an en
be the subject of a separate engagement letter that would define the scope and limits of the desired consultation services.

From: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Date: Tuesday, May 25, 2021 at 2:56 PM
To: Ken Daines <KDaines@hvjt.law>
Cc: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Subject: FW: LCvR 7(m) Confer - Preliminary Injunction Motion

Mr. Daines,

I am a trial attorney with the Department of Justice, Federal Programs Branch, and I will be defending Commerce and the Census Bureau regarding the attached case. I was forwarded your email below and would like to confer with you over your request. Are you available anytime tomorrow, May 26, between 10 am and 1 pm?

Thanks,

Jonathan

Jonathan D. Kossak
Trial Attorney | United States Department of Justice
Civil Division | Federal Programs Branch
Tel: (202) 305-0612
Email: jonathan.kossak@usdoj.gov

<image001.jpg>

This communication, along with any attachments, is covered by federal and state law governing electronic communications and may contain confidential and legally privileged information. If the reader of this message is not the intended recipient, the reader is hereby notified that any dissemination, distribution, use or copying of this message is strictly prohibited. If you have received this in error, please reply immediately to the sender and delete this message.

From: Ken Daines <KDaines@hvjt.law>
Sent: Monday, May 24, 2021 5:26:24 PM
To: General Counsel <GeneralCounsel@doc.gov>; Cannon, Michael (Federal) <MCannon@doc.gov>
Cc: Jason Torchinsky <jtorchinsky@hvjt.law>
Subject: LCvR 7(m) Confer - Preliminary Injunction Motion

Counsel:

Attached is a copy of the Complaint that our client, Fair Lines America Foundation, Inc., has filed against the Department of Commerce and the Census Bureau in the U.S. District Court for the District of Columbia.

Pursuant to DDC LCvR 7(m), we intend to file a Preliminary Injunction motion with the district court related to the claims outlined in the Complaint. Please let us know before Wednesday May 26th at 4 pm if you consent to the relief sought.

Thank you,

Ken Daines
<image002.jpg>
Office: (540) 341-8808

WWW.HVJT.LAW

PRIVILEGED AND CONFIDENTIAL

This communication and any accompanying documents are confidential and privileged. They are intended for the sole use of the addressee. If you receive this transmission in error, you are advised that any disclosure, copying, distribution, or the taking of any action in reliance upon this communication is strictly prohibited. Moreover, any such disclosure shall not compromise or waive the attorney-client, accountant-client, or other privileges as to this communication or otherwise. If you have received this communication in error, please contact me at the above email address. Thank you.

DISCLAIMER

Any accounting, business or tax advice contained in this communication, including attachments and enclosures, is not intended as a thorough, in-depth analysis of specific issues, nor a substitute for a formal opinion, nor is it sufficient to avoid tax-related penalties. If desired, Holtzman Vogel Josefiak Torchinsky, PLLC would be pleased to perform the requisite research and provide you with a detailed written analysis. Such an engagement may be the subject of a separate engagement letter that would define the scope and limits of the desired consultation services.

<2021.07.06, Second Interim Production, Fair Lines v. Commerce, No. 21-01361 (DDC), Part I.pdf>
<2021.07.06, Second Interim Production, Fair Lines v. Commerce, No. 21-01361 (DDC), Part II.pdf>

EXHIBIT 4

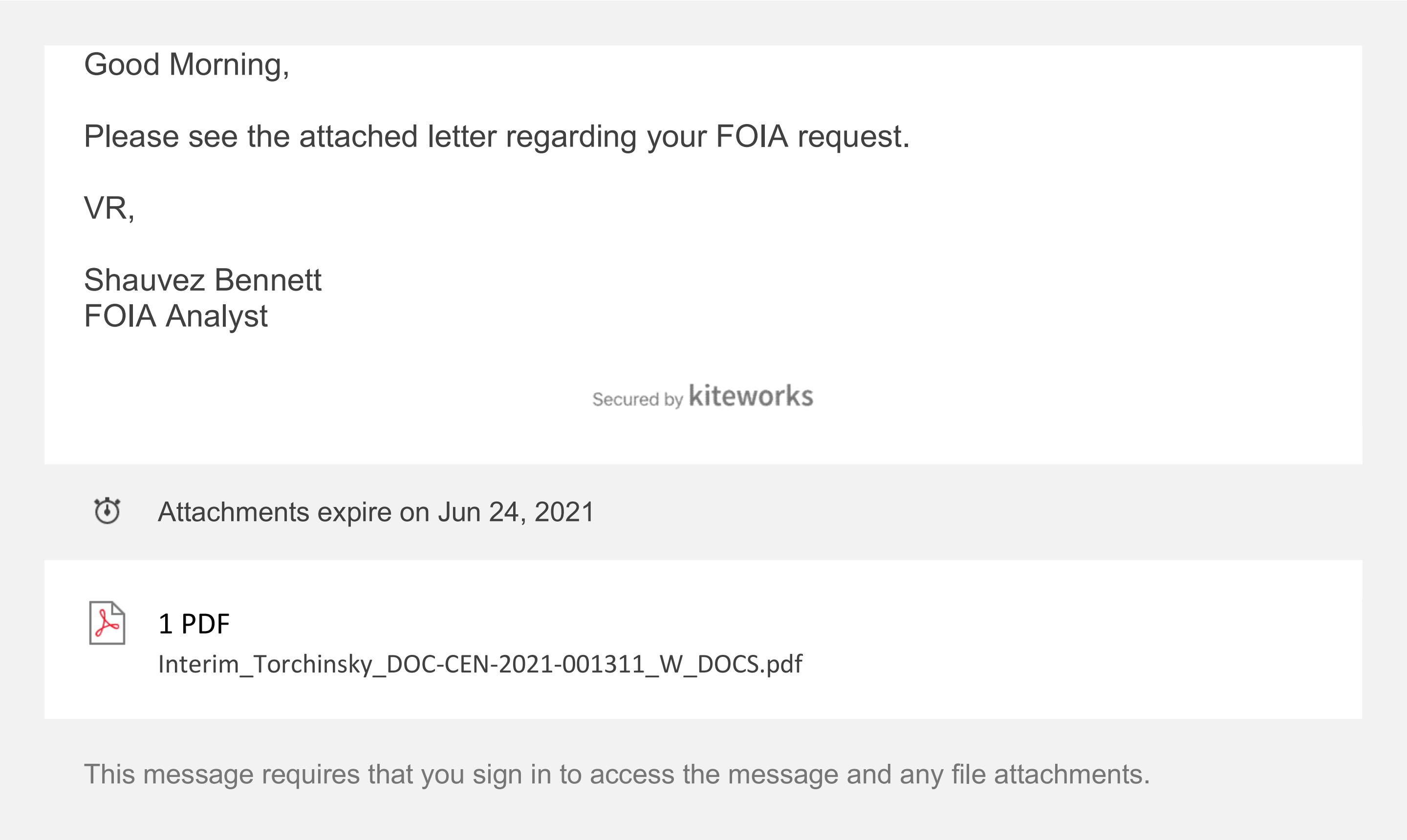
May 24 Census Bureau Determination Letter and Production

From: securefilecollaboration@doc.gov <securefilecollaboration@doc.gov>
Sent: Tuesday, May 25, 2021 9:31 AM
To: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Subject: Torchinsky_DOC-CEN-2021-001311

SECURE FILE COLLABORATION

Bennett, Shauvez sent you a secure message

Access message



Good Morning,

Please see the attached letter regarding your FOIA request.

VR,

Shauvez Bennett
FOIA Analyst

Attachments expire on Jun 24, 2021

1 PDF
Interim_Torchinsky_DOC-CEN-2021-001311_W_DOCS.pdf

This message requires that you sign in to access the message and any file attachments.

UNITED STATES DEPARTMENT OF COMMERCE
U.S. Census Bureau
Washington, DC 20233-0001

May 24, 2021

Mr. Jason Torchinsky
Fair Lines American Foundation, Inc.
2308 Mount Vernon Ave., Suite 716
Alexandria, VA 22301
jtorchinsky@hvjt.law

Dear Mr. Torchinsky:

This letter is in response to your Freedom of Information Act (FOIA), Title 5, United States Code, Section 552, request dated March 31, 2021, to the U.S. Census Bureau's FOIA Office. We received your request in this office on April 7, 2021. We have assigned to it tracking number DOC-CEN-2021-001311 and are responding under the FOIA to your request for all summaries, tabulations, and other statistical materials derived from, summarizing, and/or otherwise relating to the original underlying group quarters population data for Census Day, April 1, 2020, received in response to the Census Bureau's 2020 Group Quarters Enumeration questionnaire regarding institutional living facilities or other housing facilities.

Enclosed are 988 pages responsive to your request with withholding determinations noted. We withheld portions of the record pursuant FOIA Exemptions 3 and 5, Title 5, United States Code, Sections 552(b)(3) and (b)(5). FOIA Exemption 3 exempts from disclosure information made confidential by statute. Here, information withheld under Exemption 3 is protected by Title 13, United States Code, Section 9, which requires that census records be used solely for statistical purposes and makes these records confidential. FOIA Exemption 5 allows for the withholding of inter-agency or intra-agency memorandums or letters which would not be available by law to a party other than an agency in litigation with the agency." Here, information withheld under Exemption 5 contained information exempt from disclosure due to the deliberate process privilege, attorney client privilege, and the attorney work product privilege.

Based on the above information, this constitutes a partial denial of your request. You have the right to appeal this partial denial of the FOIA request. An appeal must be received within 90 calendar days of the date of this response letter. Address your appeal to the following office:

Assistant General Counsel for Employment, Litigation and Information
Room 5896
U.S. Department of Commerce,
14th and Constitution Avenue, N.W.
Washington, D.C. 20230

An appeal may also be sent by e-mail to FOIAAppeals@doc.gov, or by FOIAonline, if you have an account in FOIAonline, at https://foiaonline.regulations.gov/foia/action/public/home#. The appeal should include a copy of the original request and initial denial, if any. All appeals should include a statement of the reasons why the records requested should be made available and why the adverse determination was in error. The appeal letter, the envelope and the e-mail subject line should be clearly marked "Freedom of Information Act Appeal."

The e-mail and FOIAonline are monitored only on working days during normal business hours (8:30 a.m. to 5:00 p.m., Eastern Time, Monday through Friday). FOIA appeals posted to the e-mail box or FOIAonline after normal business hours will be deemed received on the next normal business day. If the 90th calendar day for submitting an appeal falls on a Saturday, Sunday or legal public holiday, an appeal received by 5:00 p.m., Eastern Time, the next business day will be deemed timely.

In addition, you may contact the Office of Government Information Services (OGIS) at the National Archives and Records Administration to inquire about the FOIA mediation services they offer. The contact information for OGIS is as follows:

Office of Government Information Services
National Archives and Records Administration
8601 Adelphi Road-OGIS
College Park, Maryland 20740-6001
e-mail at ogis@nara.gov
telephone at 202-741-5770; toll free at 1 877-684-6448
facsimile at 202-741-5769

Please contact Shauvez Bennett or Sarabeth Rodriguez of my staff, by telephone at 301-763-2127 or by email at census.efoia@census.gov if you have any questions regarding your request.

Sincerely,

Vernon Curry

Vernon E. Curry, PMP, CIPP/G
Freedom of Information Act/Privacy Act Officer
Chief, Freedom of Information Act Office

Enclosure

Enclosure

2020 Census Operational Delivery 8: Late Group Quarters Enumeration (GQE)

Thursday: October 8, 2020

Presented by: Dora Durante and Belkines Arenas Germosen

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

Pre-decisional - Internal Only - Not for Public Distribution.

DRB Approval Number: CBDRB-FY21-DSEP-002

Purpose: 2020 Census Late Group Quarters Enumeration (GQE)

- The Late GQE operation is a data collection operation established to collect respondent data for cases that met a certain criteria based on the 2020 Census Count Review Event 2 (CRO 2) operation. Cases were being conducted to provide an opportunity to obtain data for cases where FSCPE members as part of Count Review Event 2, were able to obtain additional contact information for cases marked with an outcome code of D-1 (unable to locate in block) for the following GQ type codes:

GQ Code and Description	
103 – State Prisons	501 - College/University Student Housing (owned/leased/managed by a college/university)
104 – Local Jails and Other Municipal Confinement Facilities	501 - College/University Student Housing (owned/leased/managed by a college/university)
105 - Correctional Residential Facilities	901 - Workers' Group Living Quarters and Job Corps Centers
301 - Nursing Facilities/Skilled-Nursing Facilities	999 - Unassigned or Unknown Type
601 – Military Installations	

- The Late GQE operation will also provide an opportunity for GQ administrators who missed out on the opportunity to provide respondent data for their residents prior to the end of the 2020 Census GQE operation.
- Initial plans were to have ACO staff visit identified GQ locations to collect demographic data from GQ administrators. Plans were to use NRFU enumerators who would have still been in the field during the original planned dates. Reinterview was scheduled to end one week after the completion of the data collection by enumerators.

2020 Census Late Group Quarters Enumeration (GQE) – Key Operational Activities

Timing

- Original Schedule (Pre-Pandemic):
  - Late GQE: July 1 – 29, 2020
- 1st Post – Pandemic Schedule
  - Late GQE: September 28 – Oct 23 (Reinterview: Oct 31, 2020)
- Current Schedule
  - October 1 – October 23, 2020

Pre-decisional - Internal Only - Not for Public Distribution.

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Late Group Quarters Enumeration (GQE) – Key Operational Activities

Current Activities:

- Headquarters staff across divisions are calling administrators of identified GQ locations to collect demographic data or pop count as a last resort.
  - No fieldwork
  - No reinterview
- Collection of complete resident data will help increase data quality and reduce imputation.
  - GQ administrators will receive Paper Response Data Collection (PRDC/paper listing) template to provide resident data via Kiteworks.
- Pop count alone will be a last resort.
  - Pop count data will be captured on paper listings (person 1, person 2, etc. to pop count of GQ).
- The Military Branch will conduct an outreach to Military Installations that were not enumerated during the 2020 Census GQE to verify final status (refusal, vacant, unable to locate in block, etc.) and if possible collect demographic data or pop count.

Pre-decisional - Internal Only - Not for Public Distribution.

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Late Group Quarters Enumeration (GQE) – Key Operational Activities

Current Activities continued:

HQ Staff will*:*

- Enter POP count in Max POP field in the Group Quarters Production Control System (GQPCS)
- Create Paper Listing (Excel form) for each case where only Pop count is received and that will include Person one -Pxx (number of persons in the GQ)
- For cases where a GQ admin is willing to supply complete/partial demographic data, request email address to send Kiteworks instructions.
- Place completed Paper Listings on the NPC_GQDCMD_Share (\\npc083apps) (flow basis) for retrieval and data capture using a specific naming convention: GQLATE+ 12 digit GQ ID+ YYYYMMDD
- FOCS admin will re-open cases based on GQ ID and generate completion events

For Adds (Cases not already in the GQ workload)

- HQ staff will share address along with contact information for GQPCS (cases not already in the GQ workload)
- GQPCS creates Add Case and input Actual Census Day Pop as Max Pop
- Cases will go from GQPCS > SOCS > FOCS

Pre-decisional - Internal Only - Not for Public Distribution.

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census: Late Group Quarters Enumeration – Workload Sources

Total Workload and Sources:

- Total Workload:
  1. Count Review Event 2 Results: 497
  2. Cases with paper listings or pop count Received from Group Quarters Administrators via DCMD Group Quarters eResponse email account after August 26, 2020. These included:
     a) Cases that were already in the 2020 Census GQE Workload 3608: 34
     b) Cases not in the original 2020 Census GQE workload (Adds): 47
  3. Military Cases Revisited via Phone Calls:
     a) Outcome codes of vacant, refusal, and cannot locate in block: 1972
     b) Count Review Event 2: 2

Pre-decisional - Internal Only - Not for Public Distribution.

Source: CDL, UTS

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census OD 8: Late GQE Workload Breakdown as a Result of CRO2

GQ Type Codes	Counts	Percentage of Workload
103 - State Prisons	56	11.3%
104 - Local Jails and Other Municipal Confinement Facilities	11	2.2%
105 - Correctional Residential Facilities	1	0.2%
301 - Nursing Facilities/Skilled-Nursing Facilities	123	24.7%
501 - College/University Student Housing (owned/leased/managed by a college/university)	174	35.0%
502 - College/University Student Housing (owned/leased/managed by a private company/agency)	9	1.8%
901 - Workers' Group Living Quarters and Job Corps Centers	119	23.9%
999 - Unassigned or Unknown Type	4	0.8%
Total	**497**	**100%**

Pre-decisional - Internal Only - Not for Public Distribution.

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census OD 8: Late GQE Progress

Outcome Metrics from CRO 2 as of 10/7	Count	Percentage of Workload
Assigned Cases (Total Workload)	**499**	**100%**
No of Cases Attempted	486	97%
Vacant Locations	50	10%
Non-Residential	6	1%
Refusals	34	7%
Duplicates	15	3%
Housing Units	5	1%
Transitory Locations	1	0.2%
Number of Paper Listings	2	0.4%
Pop Count only data	101	20%

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census OD 8: Late GQE Progress – Outreach from GQ Administrators

Cases Received From GQ Admins After August 26	Count	Cases Placed on NPC Share Drive	Percentage of Workload
Total Workload	81	34	41.98%
Already in GQ Universe	34	34	100.00%
Adds	47		

Pre-decisional - Internal Only - Not for Public Distribution.

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census OD 8: Late GQE Military Workload Breakdown and Progress

Military Workload	Counts	Percentage of Total Workload
Vacant	1704	86.32
Refusal	57	2.89
Unable to Locate In Block	211	10.69
Total	**1974**	

Military Progress as of 10/7	Counts	Percentage of Total Workload
Contacts Made	327	16.56%
Pop count received for paper listings	55	2.79%

Pre-decisional - Internal Only - Not for Public Distribution.

DRB Approval Number: CBDRB-FY21-DSEP-002

Backup Slides

Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

United States Census 2020

More information on the 2020 Census: http://www.census.gov/2020Census

American Community Survey

More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

instagram.com/uscensusbureau

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Operational Delivery 8: Late Group Quarters Enumeration (GQE)

Thursday: October 8, 2020

Presented by: Dora Durante and Belkines Arenas Germosen

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

Pre-decisional - Internal Only - Not for Public Distribution.

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

Purpose: 2020 Census Late Group Quarters Enumeration (GQE)

- The Late GQE operation is a data collection operation established to collect respondent data for cases that met a certain criteria based on the 2020 Census Count Review Event 2 (CRO 2) operation. Cases were being conducted to provide an opportunity to obtain data for cases where FSCPE members as part of Count Review Event 2, were able to obtain additional contact information for cases marked with an outcome code of D-1 (unable to locate in block) for the following GQ type codes:

GQ Code and Description	
103 – State Prisons	501 - College/University Student Housing (owned/leased/managed by a college/university)
104 – Local Jails and Other Municipal Confinement Facilities	501 - College/University Student Housing (owned/leased/managed by a college/university)
105 - Correctional Residential Facilities	901 - Workers' Group Living Quarters and Job Corps Centers
301 - Nursing Facilities/Skilled-Nursing Facilities	999 - Unassigned or Unknown Type
601 – Military Installations	

- The Late GQE operation will also provide an opportunity for GQ administrators who missed out on the opportunity to provide respondent data for their residents prior to the end of the 2020 Census GQE operation.
- Initial plans were to have ACO staff visit identified GQ locations to collect demographic data from GQ administrators. Plans were to use NRFU enumerators who would have still been in the field during the original planned dates. Reinterview was scheduled to end one week after the completion of the data collection by enumerators.

Pre-decisional - Internal Only - Not for Public Distribution.

2020 Census Late Group Quarters Enumeration (GQE) – Key Operational Activities

Timing

- Original Schedule (Pre-Pandemic):
  - Late GQE: July 1 – 29, 2020

- 1st Post – Pandemic Schedule
  - Late GQE: September 28 – Oct 23 (Reinterview: Oct 31, 2020)

- Current Schedule
  - October 1 – October 23, 2020

Pre-decisional - Internal Only - Not for Public Distribution.

2020 Census Late Group Quarters Enumeration (GQE) – Key Operational Activities

Current Activities:

- Headquarters staff across divisions are calling administrators of identified GQ locations to collect demographic data or pop count as a last resort.
  - No fieldwork
  - No reinterview
- Collection of complete resident data will help increase data quality and reduce imputation.
  - GQ administrators will receive Paper Response Data Collection (PRDC/paper listing) template to provide resident data via Kiteworks.
- Pop count alone will be a last resort.
  - Pop count data will be captured on paper listings (person 1, person 2, etc. to pop count of GQ).

- The Military Branch will conduct an outreach to Military Installations that were not enumerated during the 2020 Census GQE to verify final status (refusal, vacant, unable to locate in block, etc.) and if possible collect demographic data or pop count.

2020 Census Late Group Quarters Enumeration (GQE) – Key Operational Activities

Current Activities continued:

HQ Staff will*:*

- Enter POP count in Max POP field in the Group Quarters Production Control System (GQPCS)
- Create Paper Listing (Excel form) for each case where only Pop count is received and that will include Person one -Pxx (number of persons in the GQ)
- For cases where a GQ admin is willing to supply complete/partial demographic data, request email address to send Kiteworks instructions.
- Place completed Paper Listings on the NPC_GQDCMD_Share (\\npc083apps) (flow basis) for retrieval and data capture using a specific naming convention: GQLATE+ 12 digit GQ ID+ YYYYMMDD
- FOCS admin will re-open cases based on GQ ID and generate completion events

For Adds (Cases not already in the GQ workload)

- HQ staff will share address along with contact information for GQPCS (cases not already in the GQ workload)
- GQPCS creates Add Case and input Actual Census Day Pop as Max Pop
- Cases will go from GQPCS > SOCS > FOCS

Pre-decisional - Internal Only - Not for Public Distribution.

2020 Census: Late Group Quarters Enumeration – Workload Sources

Total Workload and Sources:

- Total Workload:
  1. Count Review Event 2 Results: 497
  2. Cases with paper listings or pop count Received from Group Quarters Administrators via DCMD Group Quarters eResponse email account after August 26, 2020. These included:
     a) Cases that were already in the 2020 Census GQE Workload 3608: 34
     b) Cases not in the original 2020 Census GQE workload (Adds): 47
  3. Military Cases Revisited via Phone Calls:
     a) Outcome codes of vacant, refusal, and cannot locate in block: 1972
     b) Count Review Event 2: 2

Pre-decisional - Internal Only - Not for Public Distribution.

Source: CDL, UTS

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census OD 8: Late GQE Workload Breakdown as a Result of CRO2

GQ Type Codes	Counts	Percentage of Workload
103 - State Prisons	56	11.3%
104 - Local Jails and Other Municipal Confinement Facilities	11	2.2%
105 - Correctional Residential Facilities	1	0.2%
301 - Nursing Facilities/Skilled-Nursing Facilities	123	24.7%
501 - College/University Student Housing (owned/leased/managed by a college/university)	174	35.0%
502 - College/University Student Housing (owned/leased/managed by a private company/agency)	9	1.8%
901 - Workers' Group Living Quarters and Job Corps Centers	119	23.9%
999 - Unassigned or Unknown Type	4	0.8%
Total	**497**	**100%**

Pre-decisional - Internal Only - Not for Public Distribution.

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census OD 8: Late GQE Progress

Outcome Metrics from CRO 2 as of 10/7	Count	Percentage of Workload
Assigned Cases (Total Workload)	**497**	**100%**
No of Cases Attempted	484	97%
Vacant Locations	50	10%
Non-Residential	6	1%
Refusals	34	7%
Duplicates	15	3%
Housing Units	5	1%
Transitory Locations	1	0.2%
Number of Paper Listings	2	0.4%
Pop Count only data	101	20%

2020 Census OD 8: Late GQE Progress – Outreach from GQ Administrators

Cases Received From GQ Admins After August 26	Count	Cases Placed on NPC Share Drive	Percentage of Workload
Total Workload	81	34	41.98%
Already in GQ Universe	34	34	100.00%
Adds	47		

Pre-decisional - Internal Only - Not for Public Distribution.

2020 Census OD 8: Late GQE Military Workload Breakdown and Progress

Military Workload	Counts	Percentage of Total Workload
Vacant	1704	86.32
Refusal	57	2.89
Unable to Locate In Block	211	10.69
Total	**1974**	

Military Progress as of 10/7	Counts	Percentage of Total Workload
Contacts Made	327	16.56%
Pop count received for paper listings	55	2.79%

Backup Slides

Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

More information on the 2020 Census: http://www.census.gov/2020Census

More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

instagram.com/uscensusbureau

2020 Census Operational Delivery 8: Group Quarters Update

Focus: Group Quarters Enumeration and Maritime/Military Vessel Enumeration

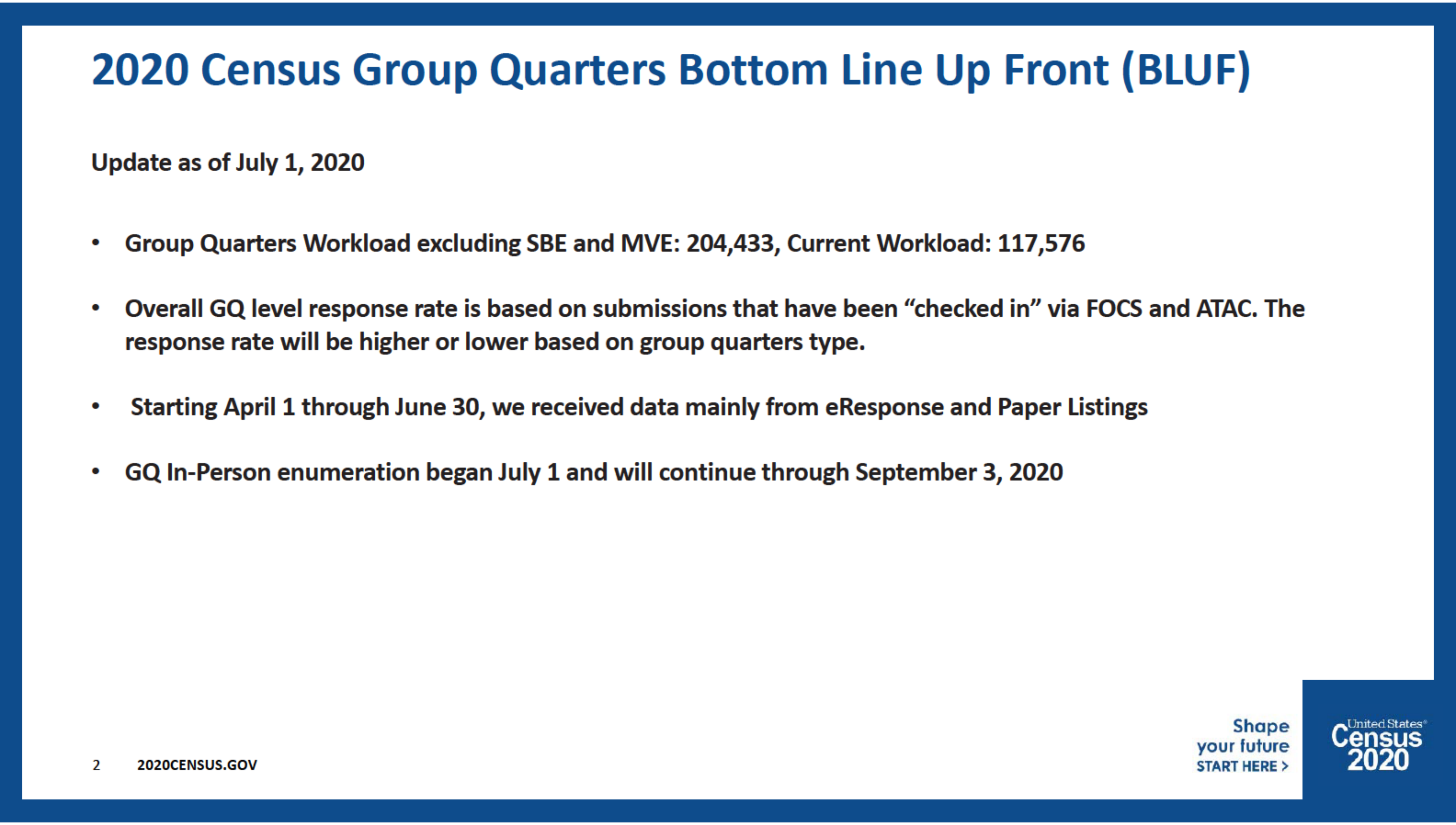
Thursday: July 2, 2020

Presented by: Dora Durante and Crystal Miller

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

Shape
your future
START HERE >

United States®
Census
2020



2020 Census Group Quarters Bottom Line Up Front (BLUF)

Update as of July 1, 2020

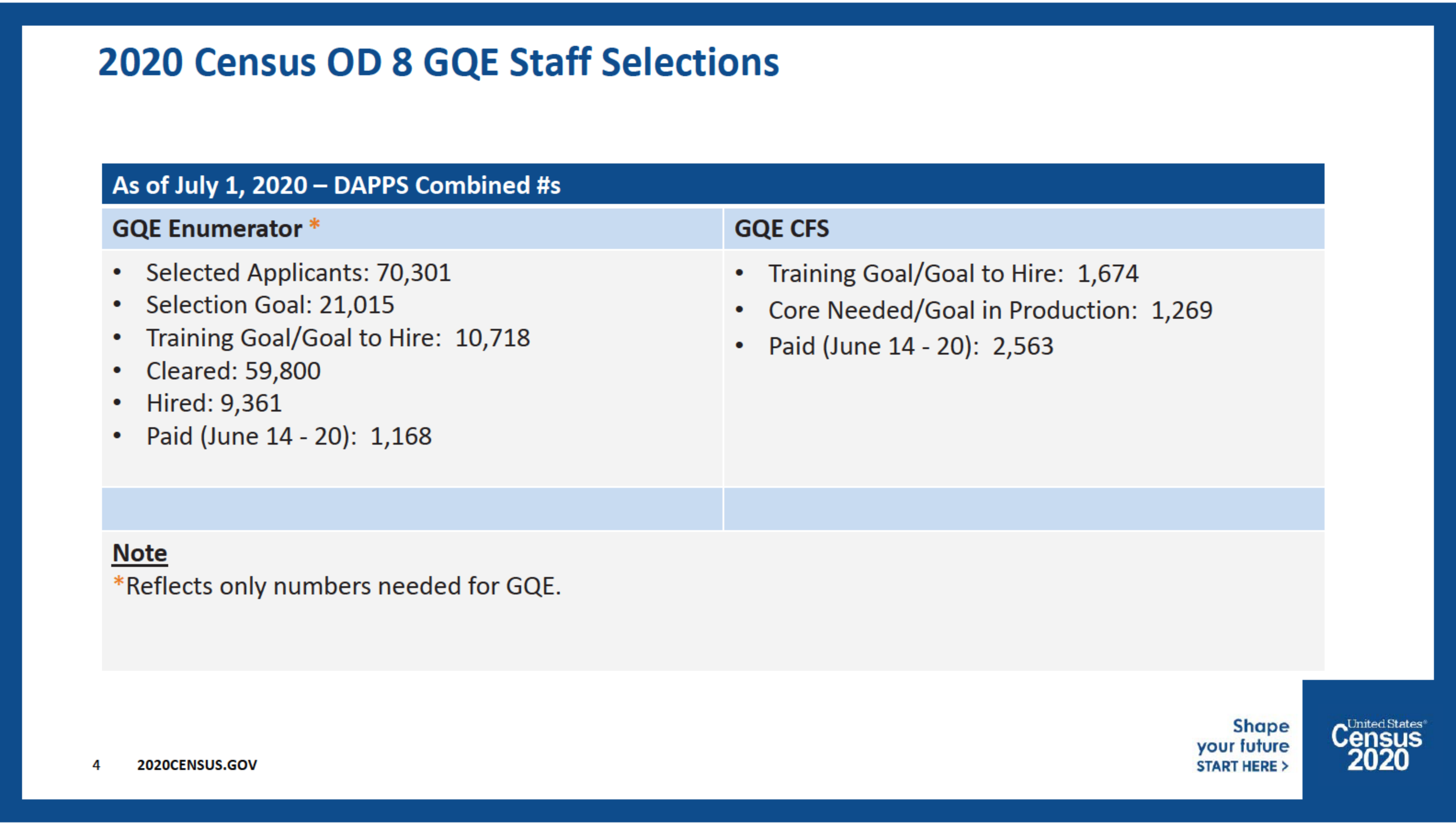
- **Group Quarters Workload excluding SBE and MVE: 204,433, Current Workload: 117,576**
- **Overall GQ level response rate is based on submissions that have been "checked in" via FOCS and ATAC. The response rate will be higher or lower based on group quarters type.**
- **Starting April 1 through June 30, we received data mainly from eResponse and Paper Listings**
- **GQ In-Person enumeration began July 1 and will continue through September 3, 2020**

OD 8 GQE Milestone Conduct & Training Dates

Conduct Activity	Proposed Start	Proposed Finish
Conduct GQE eResponse Operation	**04/01/20 (A)**	**08/07/20 (P)**
Conduct GQE Operation Field	**04/20/20 (A)**	**08/26/20 (P)**
Conduct GQE In-Person Operation	**07/01/20 (A)**	08/26/20 (P)
Conduct GQ Reinterview Operation	07/02/20 (P)	09/03/20 (P)

Training Activity	Proposed Start	Proposed Finish
Conduct GQE CFS Refresher Training	**06/16/20 (A)**	**06/19/20 (A)**
Conduct GQE Clerk Refresher Training	**06/19/20 (A)**	
Conduct GQE Enumerator Training	**06/26/20 (A)**	**07/01/20 (A)**

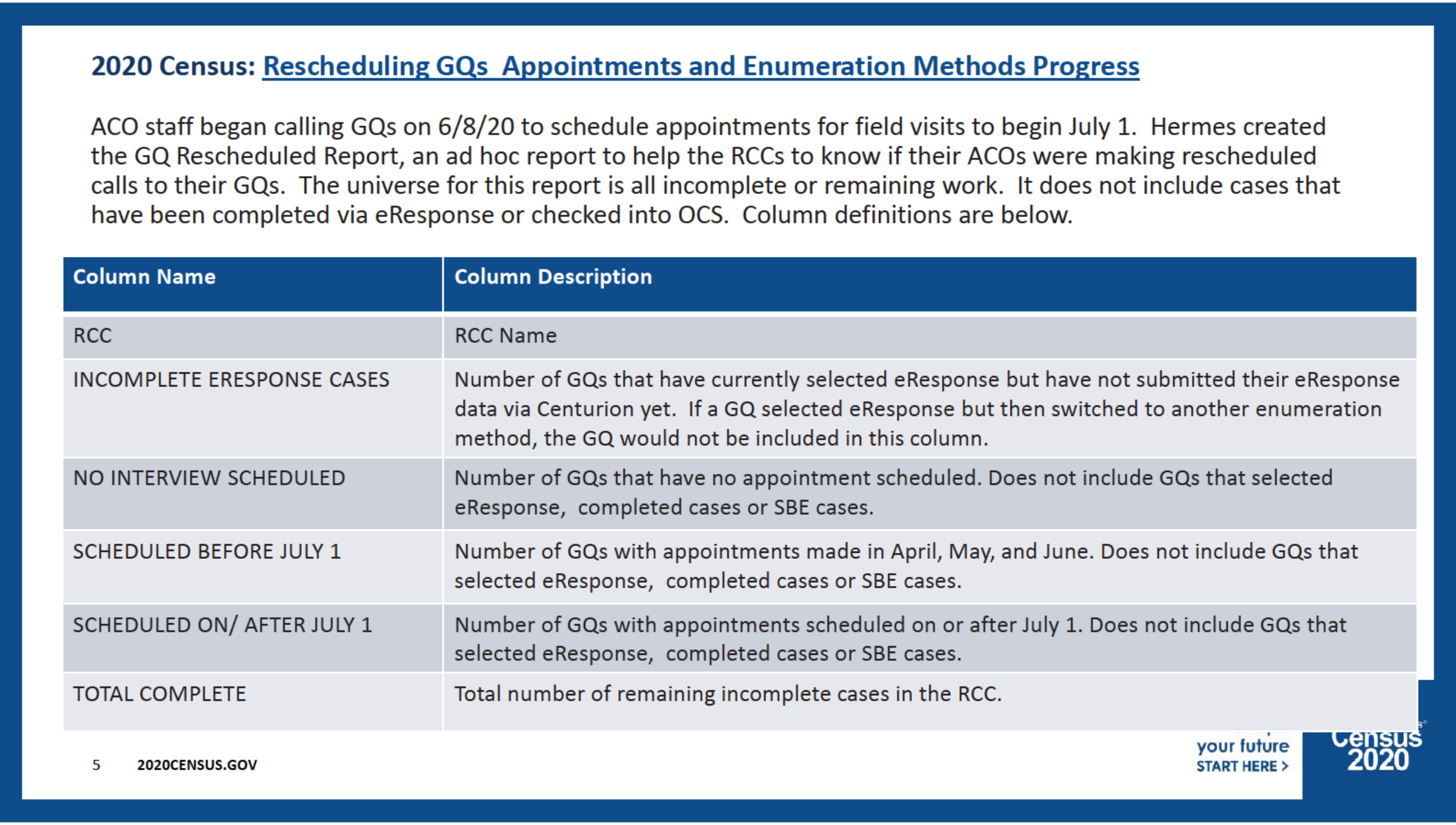
DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census OD 8 GQE Staff Selections

As of July 1, 2020 – DAPPS Combined #s	
GQE Enumerator *	**GQE CFS**
• Selected Applicants: 70,301 • Selection Goal: 21,015 • Training Goal/Goal to Hire: 10,718 • Cleared: 59,800 • Hired: 9,361 • Paid (June 14 - 20): 1,168	• Training Goal/Goal to Hire: 1,674 • Core Needed/Goal in Production: 1,269 • Paid (June 14 - 20): 2,563
Note *Reflects only numbers needed for GQE.	

DRB Approval Number: CBDRB-FY21-DSEP-002

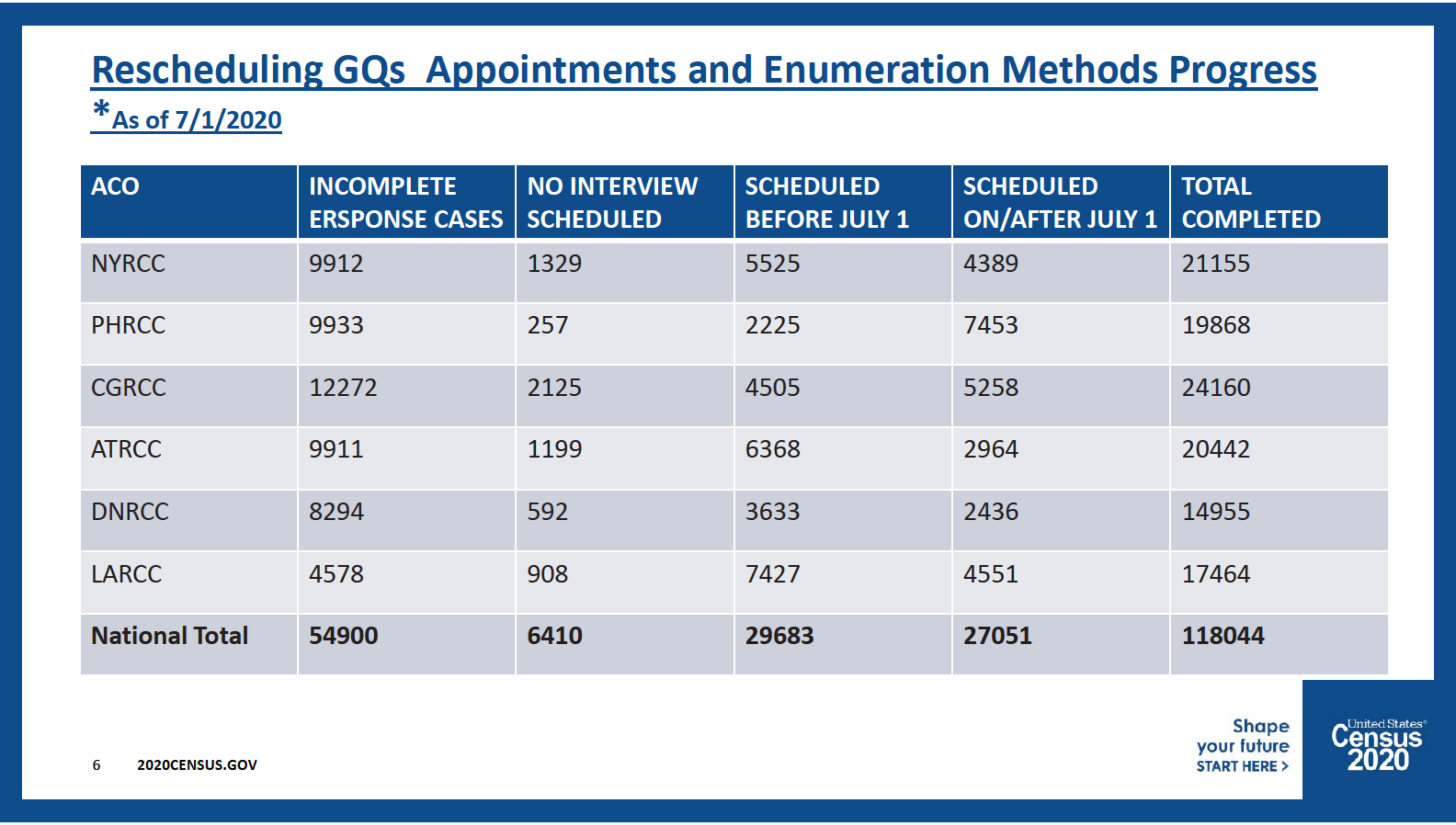


2020 Census: Rescheduling GQs Appointments and Enumeration Methods Progress

ACO staff began calling GQs on 6/8/20 to schedule appointments for field visits to begin July 1. Hermes created the GQ Rescheduled Report, an ad hoc report to help the RCCs to know if their ACOs were making rescheduled calls to their GQs. The universe for this report is all incomplete or remaining work. It does not include cases that have been completed via eResponse or checked into OCS. Column definitions are below.

Column Name	Column Description
RCC	RCC Name
INCOMPLETE ERESPONSE CASES	Number of GQs that have currently selected eResponse but have not submitted their eResponse data via Centurion yet. If a GQ selected eResponse but then switched to another enumeration method, the GQ would not be included in this column.
NO INTERVIEW SCHEDULED	Number of GQs that have no appointment scheduled. Does not include GQs that selected eResponse, completed cases or SBE cases.
SCHEDULED BEFORE JULY 1	Number of GQs with appointments made in April, May, and June. Does not include GQs that selected eResponse, completed cases or SBE cases.
SCHEDULED ON/ AFTER JULY 1	Number of GQs with appointments scheduled on or after July 1. Does not include GQs that selected eResponse, completed cases or SBE cases.
TOTAL COMPLETE	Total number of remaining incomplete cases in the RCC.

DRB Approval Number: CBDRB-FY21-DSEP-002



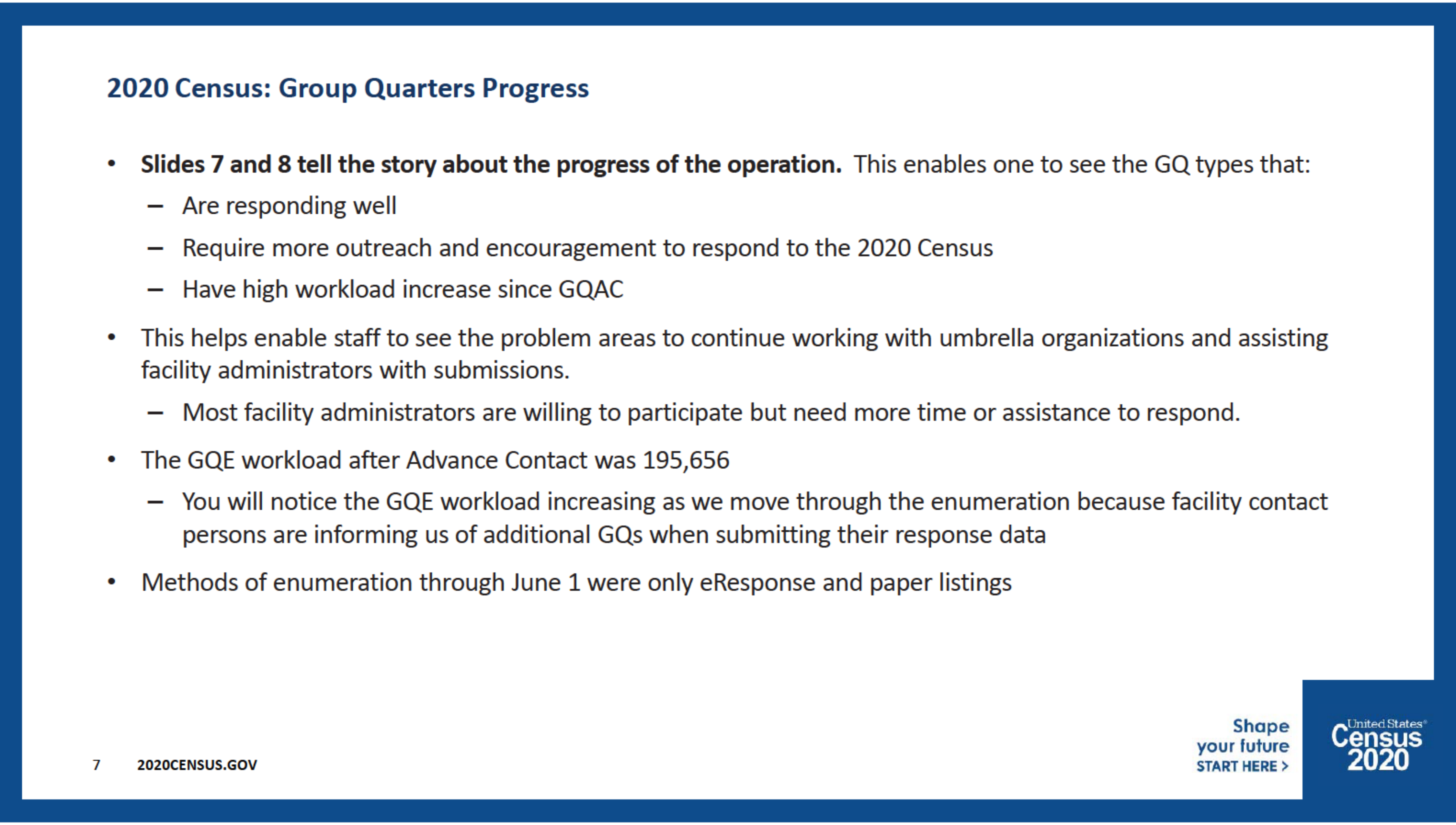
Rescheduling GQs Appointments and Enumeration Methods Progress

*As of 7/1/2020

ACO	INCOMPLETE ERSPONSE CASES	NO INTERVIEW SCHEDULED	SCHEDULED BEFORE JULY 1	SCHEDULED ON/AFTER JULY 1	TOTAL COMPLETED
NYRCC	9912	1329	5525	4389	21155
PHRCC	9933	257	2225	7453	19868
CGRCC	12272	2125	4505	5258	24160
ATRCC	9911	1199	6368	2964	20442
DNRCC	8294	592	3633	2436	14955
LARCC	4578	908	7427	4551	17464
National Total	**54900**	**6410**	**29683**	**27051**	**118044**

6 2020CENSUS.GOV

Shape
your future
START HERE >



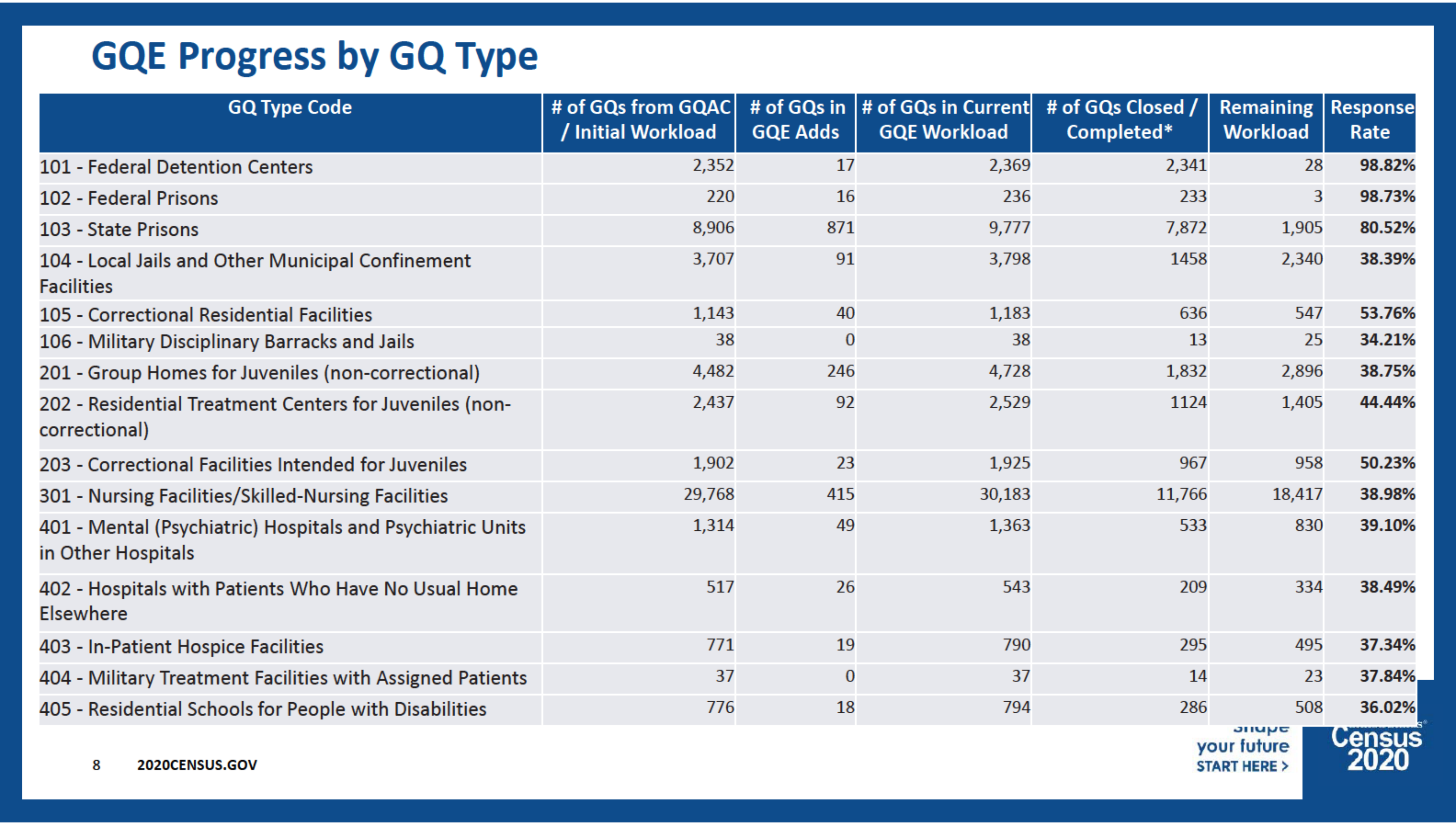
2020 Census: Group Quarters Progress

- **Slides 7 and 8 tell the story about the progress of the operation.** This enables one to see the GQ types that:
  - Are responding well
  - Require more outreach and encouragement to respond to the 2020 Census
  - Have high workload increase since GQAC
- This helps enable staff to see the problem areas to continue working with umbrella organizations and assisting facility administrators with submissions.
  - Most facility administrators are willing to participate but need more time or assistance to respond.
- The GQE workload after Advance Contact was 195,656
  - You will notice the GQE workload increasing as we move through the enumeration because facility contact persons are informing us of additional GQs when submitting their response data
- Methods of enumeration through June 1 were only eResponse and paper listings

Shape
your future
START HERE >

United States®
Census
2020

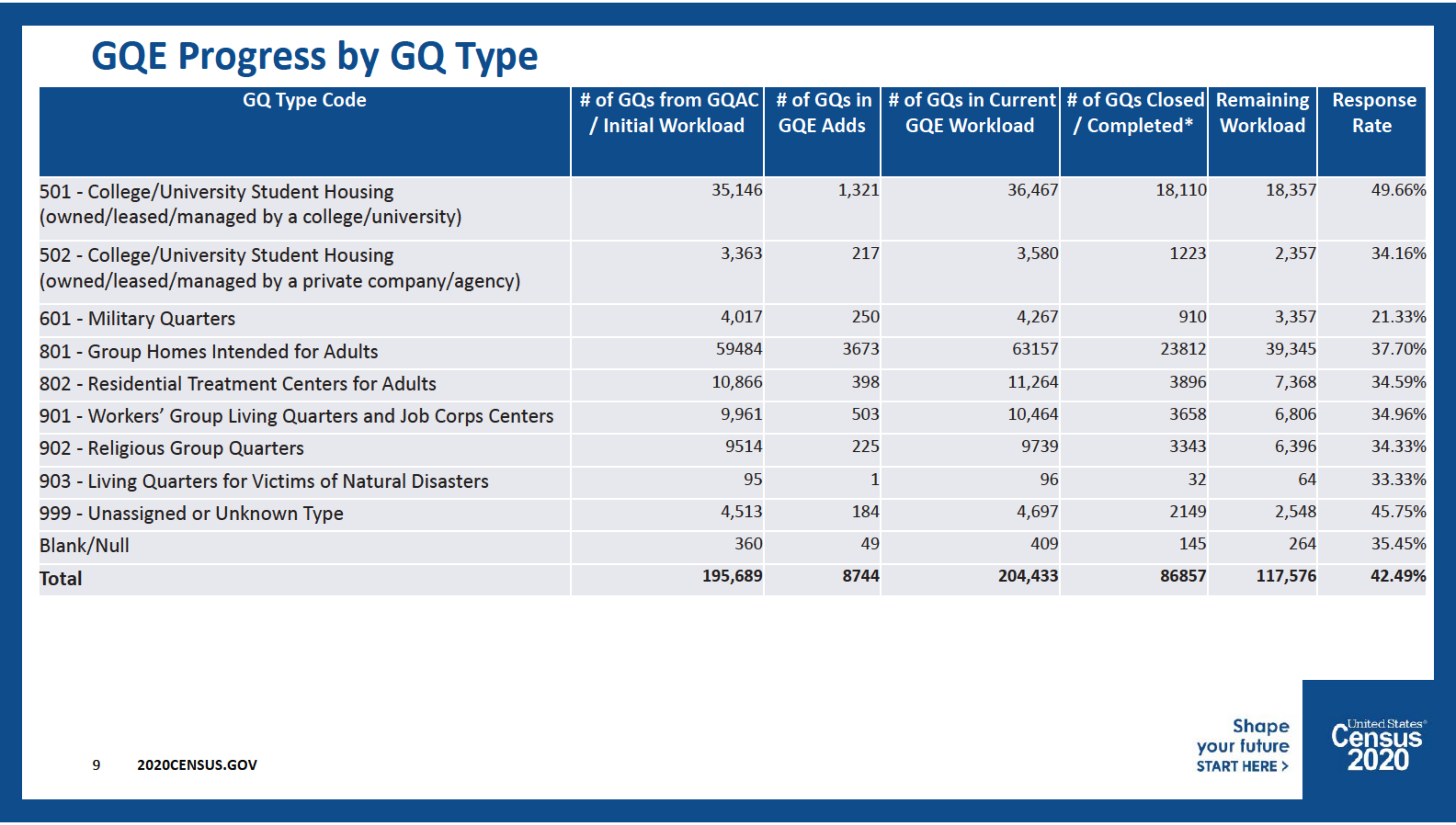
DRB Approval Number: CBDRB-FY21-DSEP-002



GQE Progress by GQ Type

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Current GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
101 - Federal Detention Centers	2,352	17	2,369	2,341	28	**98.82%**
102 - Federal Prisons	220	16	236	233	3	**98.73%**
103 - State Prisons	8,906	871	9,777	7,872	1,905	**80.52%**
104 - Local Jails and Other Municipal Confinement Facilities	3,707	91	3,798	1458	2,340	**38.39%**
105 - Correctional Residential Facilities	1,143	40	1,183	636	547	**53.76%**
106 - Military Disciplinary Barracks and Jails	38	0	38	13	25	**34.21%**
201 - Group Homes for Juveniles (non-correctional)	4,482	246	4,728	1,832	2,896	**38.75%**
202 - Residential Treatment Centers for Juveniles (non-correctional)	2,437	92	2,529	1124	1,405	**44.44%**
203 - Correctional Facilities Intended for Juveniles	1,902	23	1,925	967	958	**50.23%**
301 - Nursing Facilities/Skilled-Nursing Facilities	29,768	415	30,183	11,766	18,417	**38.98%**
401 - Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals	1,314	49	1,363	533	830	**39.10%**
402 - Hospitals with Patients Who Have No Usual Home Elsewhere	517	26	543	209	334	**38.49%**
403 - In-Patient Hospice Facilities	771	19	790	295	495	**37.34%**
404 - Military Treatment Facilities with Assigned Patients	37	0	37	14	23	**37.84%**
405 - Residential Schools for People with Disabilities	776	18	794	286	508	**36.02%**

DRB Approval Number: CBDRB-FY21-DSEP-002



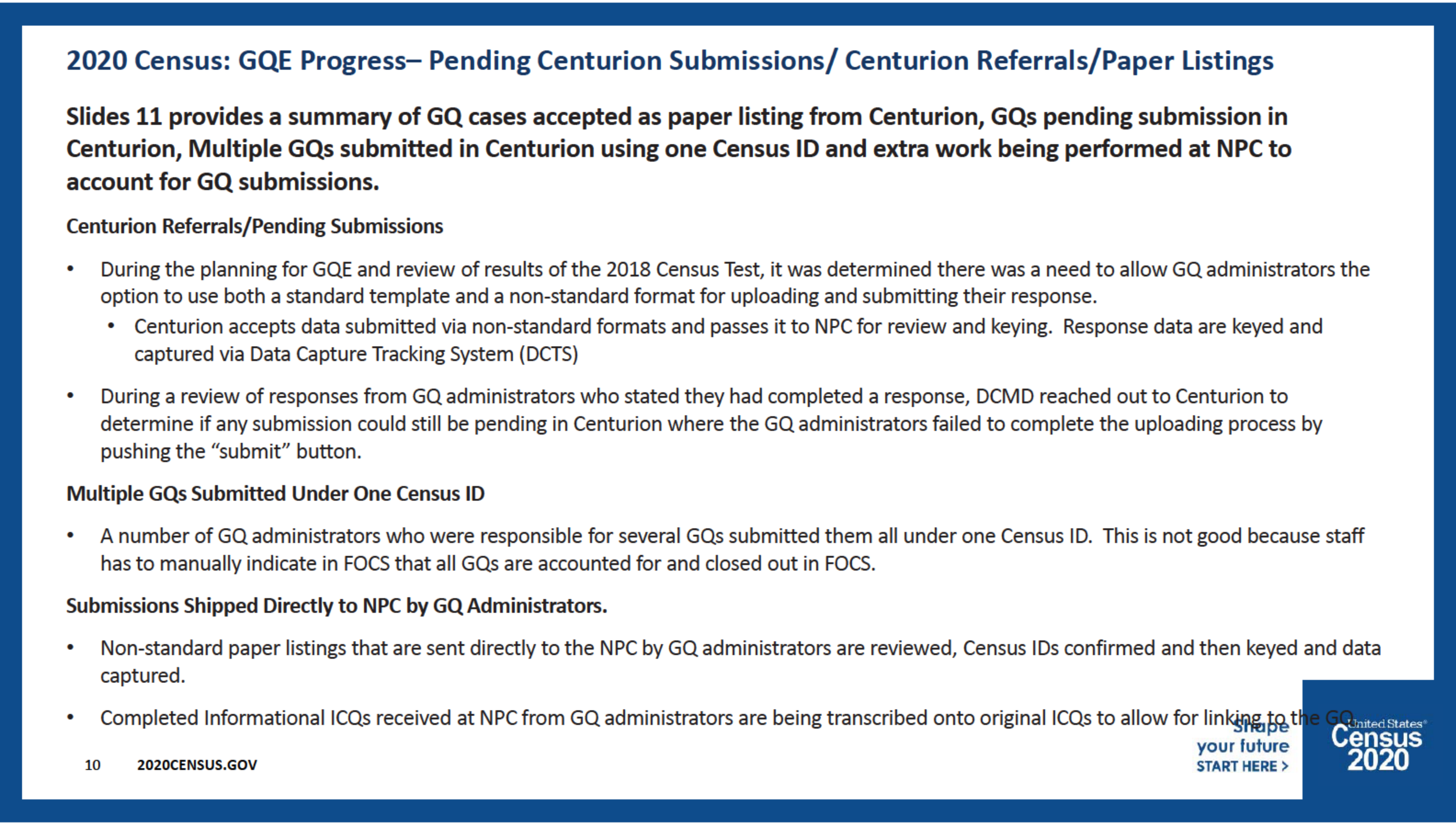
GQE Progress by GQ Type

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Current GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
501 - College/University Student Housing (owned/leased/managed by a college/university)	35,146	1,321	36,467	18,110	18,357	49.66%
502 - College/University Student Housing (owned/leased/managed by a private company/agency)	3,363	217	3,580	1223	2,357	34.16%
601 - Military Quarters	4,017	250	4,267	910	3,357	21.33%
801 - Group Homes Intended for Adults	59484	3673	63157	23812	39,345	37.70%
802 - Residential Treatment Centers for Adults	10,866	398	11,264	3896	7,368	34.59%
901 - Workers' Group Living Quarters and Job Corps Centers	9,961	503	10,464	3658	6,806	34.96%
902 - Religious Group Quarters	9514	225	9739	3343	6,396	34.33%
903 - Living Quarters for Victims of Natural Disasters	95	1	96	32	64	33.33%
999 - Unassigned or Unknown Type	4,513	184	4,697	2149	2,548	45.75%
Blank/Null	360	49	409	145	264	35.45%
Total	**195,689**	**8744**	**204,433**	**86857**	**117,576**	**42.49%**

Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census: GQE Progress– Pending Centurion Submissions/ Centurion Referrals/Paper Listings

Slides 11 provides a summary of GQ cases accepted as paper listing from Centurion, GQs pending submission in Centurion, Multiple GQs submitted in Centurion using one Census ID and extra work being performed at NPC to account for GQ submissions.

Centurion Referrals/Pending Submissions

- During the planning for GQE and review of results of the 2018 Census Test, it was determined there was a need to allow GQ administrators the option to use both a standard template and a non-standard format for uploading and submitting their response.
  - Centurion accepts data submitted via non-standard formats and passes it to NPC for review and keying. Response data are keyed and captured via Data Capture Tracking System (DCTS)
- During a review of responses from GQ administrators who stated they had completed a response, DCMD reached out to Centurion to determine if any submission could still be pending in Centurion where the GQ administrators failed to complete the uploading process by pushing the “submit” button.

Multiple GQs Submitted Under One Census ID

- A number of GQ administrators who were responsible for several GQs submitted them all under one Census ID. This is not good because staff has to manually indicate in FOCS that all GQs are accounted for and closed out in FOCS.

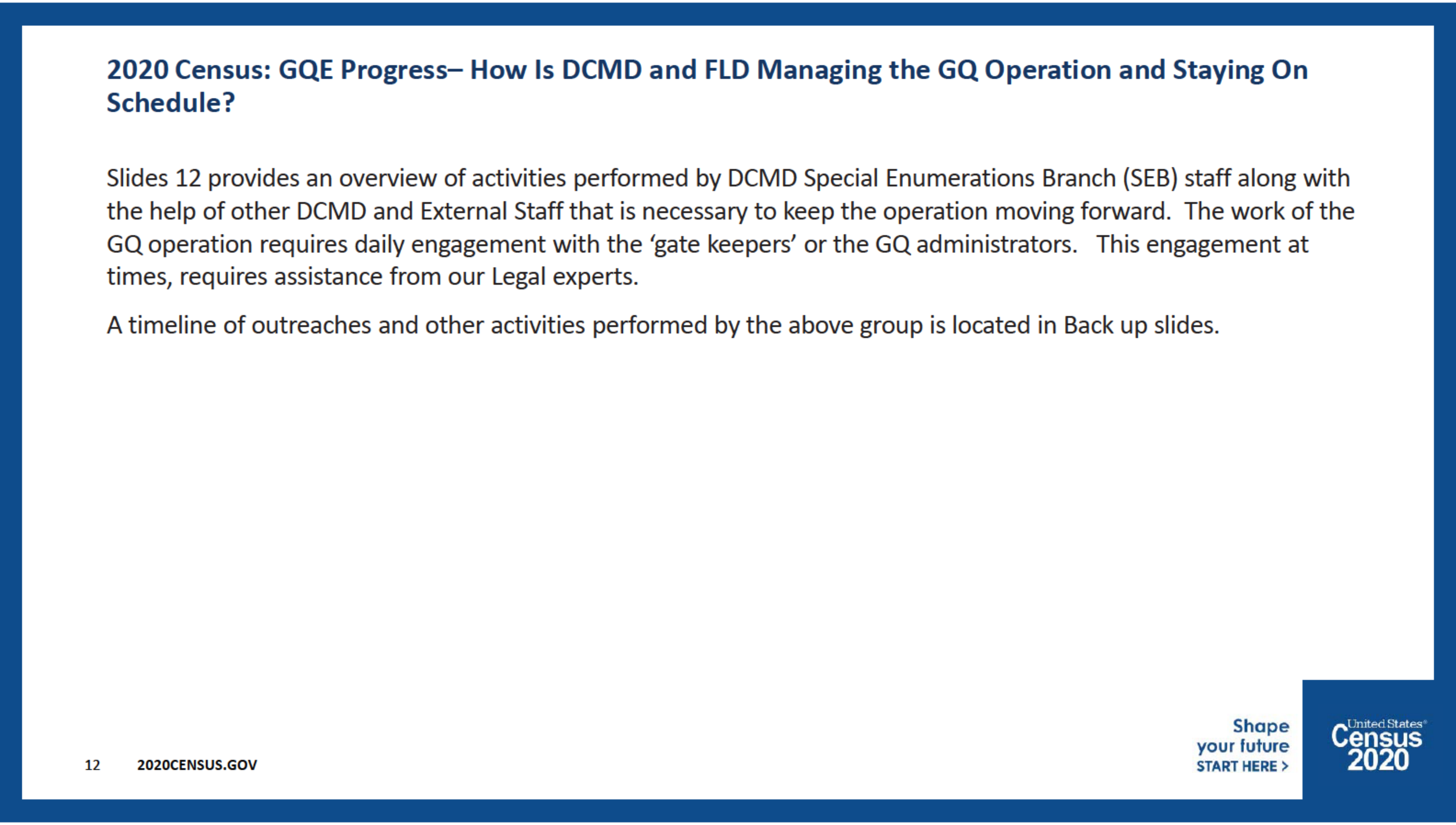
Submissions Shipped Directly to NPC by GQ Administrators.

- Non-standard paper listings that are sent directly to the NPC by GQ administrators are reviewed, Census IDs confirmed and then keyed and data captured.
- Completed Informational ICQs received at NPC from GQ administrators are being transcribed onto original ICQs to allow for linking to the GQ.

GQE Progress: Centurion Referrals /Pending Submissions/Paper Listings

	Centurion Referral Paper Listing (non-standard formats)	Submissions Remaining with Pending Status in Centurion	One GQ Census IDs with Possibly Multiple GQs Included*
Received	700	389	1167 where a diff ≥ 10

DRB Approval Number: CBDRB-FY21-DSEP-002



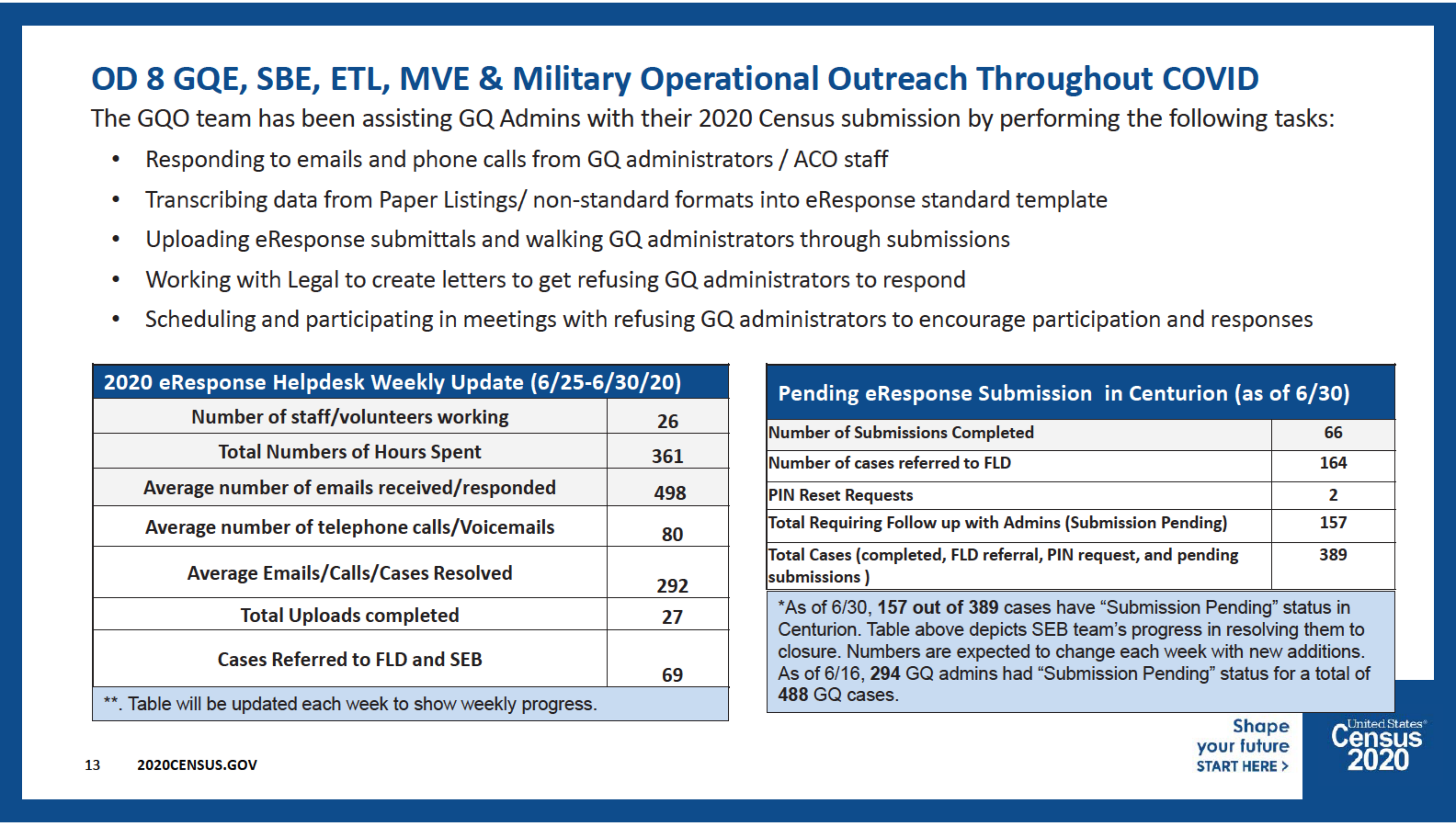
2020 Census: GQE Progress– How Is DCMD and FLD Managing the GQ Operation and Staying On Schedule?

Slides 12 provides an overview of activities performed by DCMD Special Enumerations Branch (SEB) staff along with the help of other DCMD and External Staff that is necessary to keep the operation moving forward. The work of the GQ operation requires daily engagement with the 'gate keepers' or the GQ administrators. This engagement at times, requires assistance from our Legal experts.

A timeline of outreaches and other activities performed by the above group is located in Back up slides.

Shape your future START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



OD 8 GQE, SBE, ETL, MVE & Military Operational Outreach Throughout COVID

The GQO team has been assisting GQ Admins with their 2020 Census submission by performing the following tasks:

- Responding to emails and phone calls from GQ administrators / ACO staff
- Transcribing data from Paper Listings/ non-standard formats into eResponse standard template
- Uploading eResponse submittals and walking GQ administrators through submissions
- Working with Legal to create letters to get refusing GQ administrators to respond
- Scheduling and participating in meetings with refusing GQ administrators to encourage participation and responses

2020 eResponse Helpdesk Weekly Update (6/25-6/30/20)	
Number of staff/volunteers working	26
Total Numbers of Hours Spent	361
Average number of emails received/responded	498
Average number of telephone calls/Voicemails	80
Average Emails/Calls/Cases Resolved	292
Total Uploads completed	27
Cases Referred to FLD and SEB	69

**. Table will be updated each week to show weekly progress.

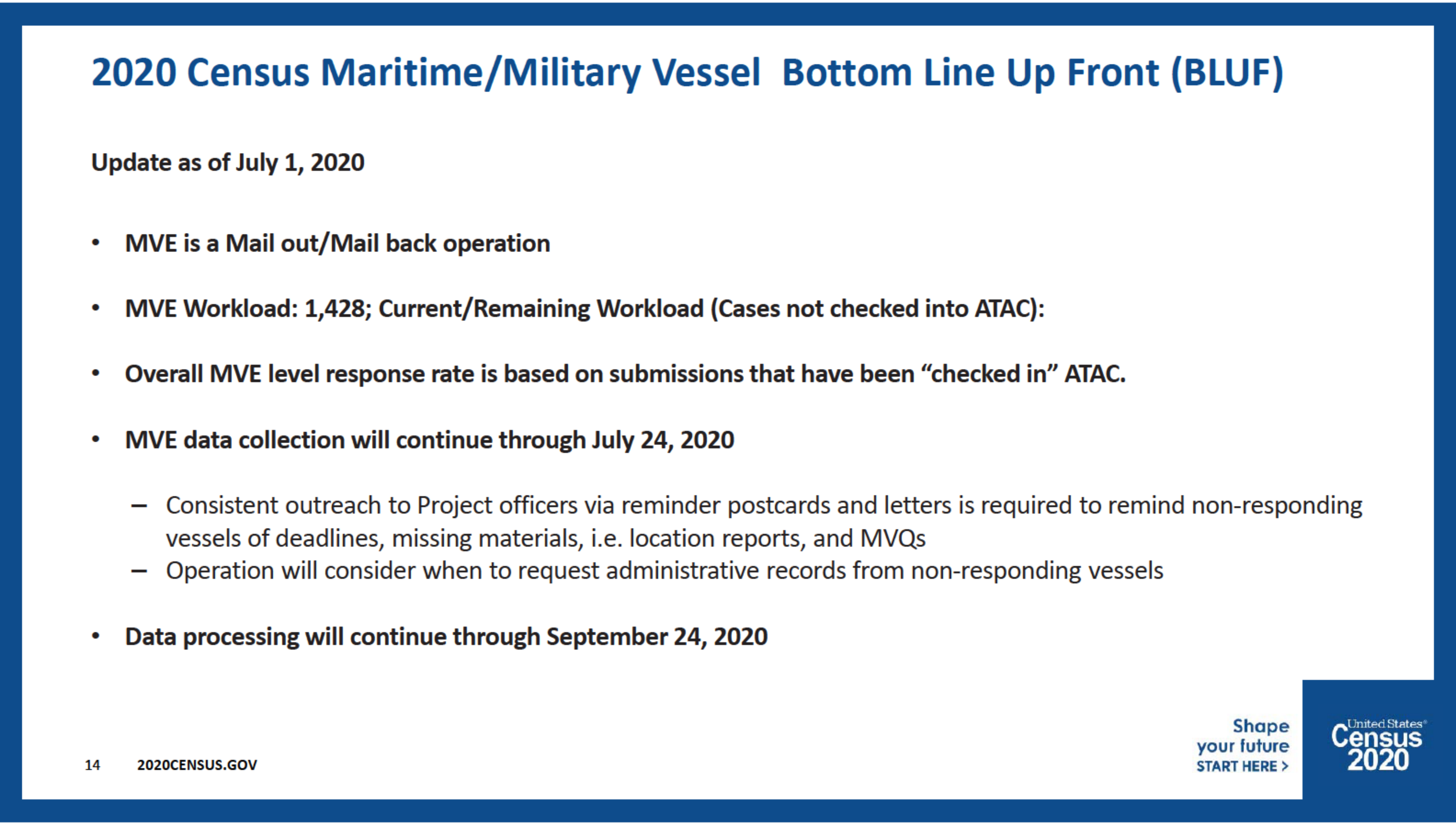
Pending eResponse Submission in Centurion (as of 6/30)	
Number of Submissions Completed	66
Number of cases referred to FLD	164
PIN Reset Requests	2
Total Requiring Follow up with Admins (Submission Pending)	157
Total Cases (completed, FLD referral, PIN request, and pending submissions)	389

*As of 6/30, **157 out of 389** cases have "Submission Pending" status in Centurion. Table above depicts SEB team's progress in resolving them to closure. Numbers are expected to change each week with new additions. As of 6/16, **294** GQ admins had "Submission Pending" status for a total of **488** GQ cases.

Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



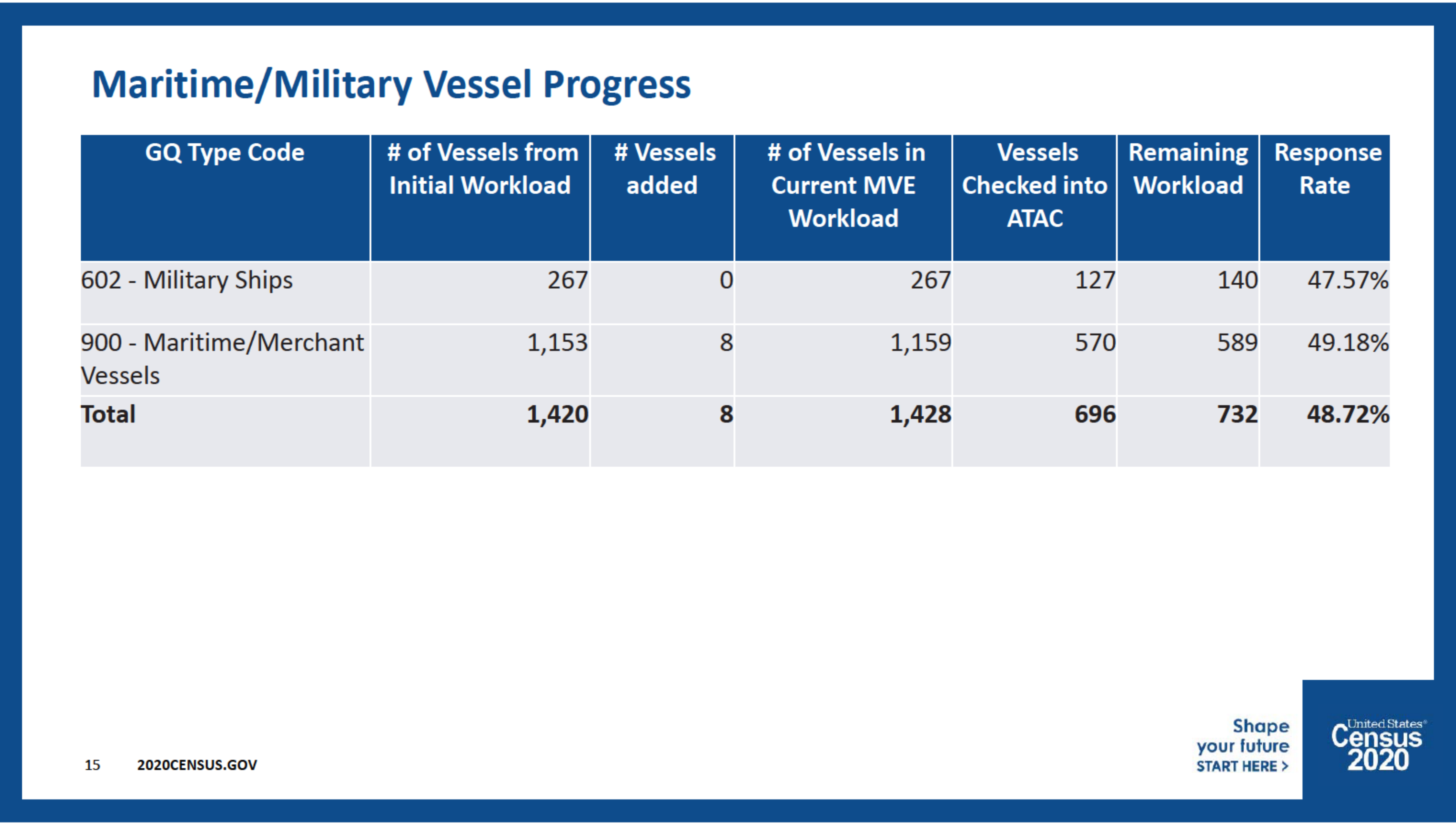
2020 Census Maritime/Military Vessel Bottom Line Up Front (BLUF)

Update as of July 1, 2020

- **MVE is a Mail out/Mail back operation**
- **MVE Workload: 1,428; Current/Remaining Workload (Cases not checked into ATAC):**
- **Overall MVE level response rate is based on submissions that have been “checked in” ATAC.**
- **MVE data collection will continue through July 24, 2020**
  - Consistent outreach to Project officers via reminder postcards and letters is required to remind non-responding vessels of deadlines, missing materials, i.e. location reports, and MVQs
  - Operation will consider when to request administrative records from non-responding vessels
- **Data processing will continue through September 24, 2020**

Shape
your future
START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002

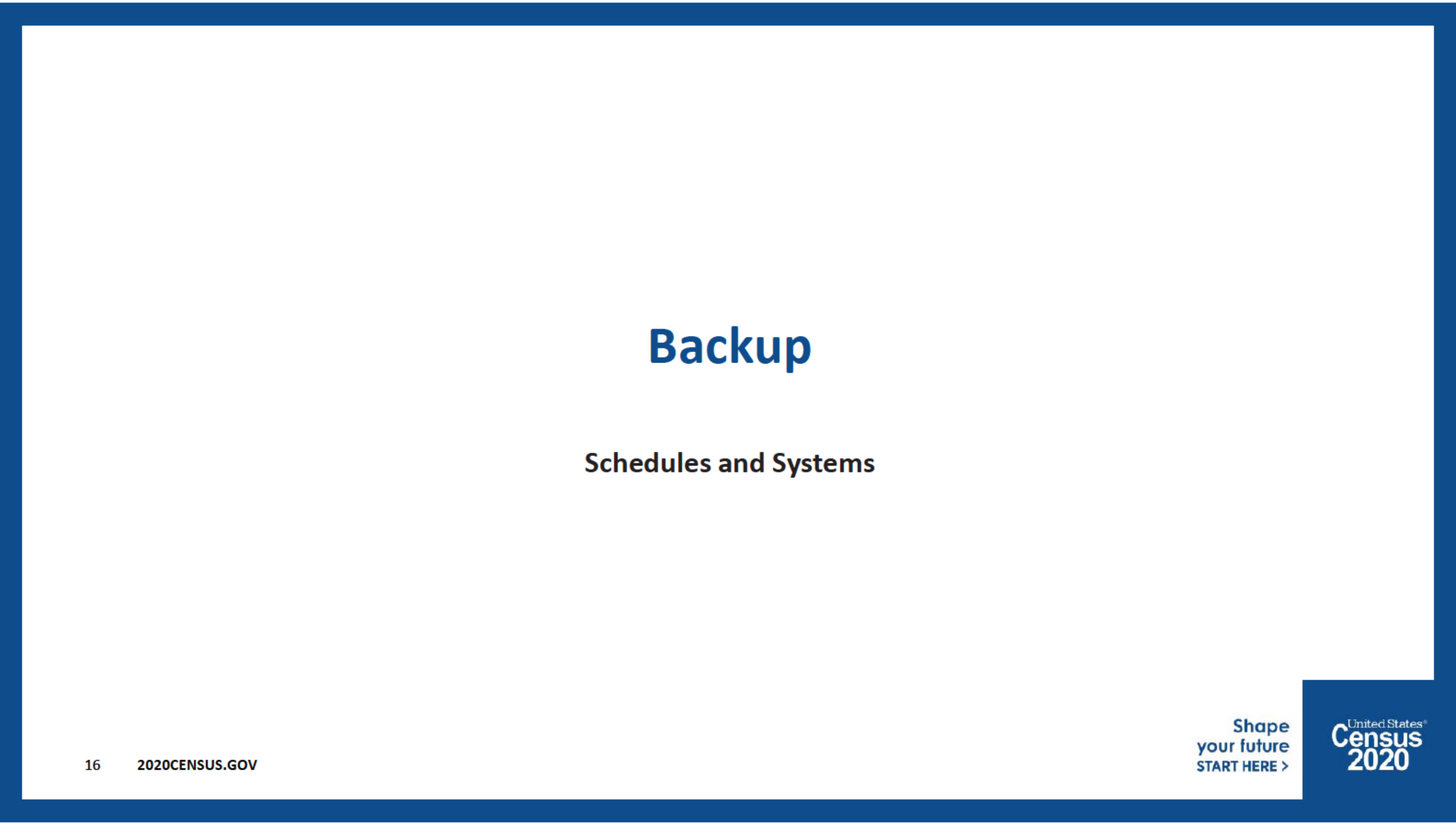


Maritime/Military Vessel Progress

GQ Type Code	# of Vessels from Initial Workload	# Vessels added	# of Vessels in Current MVE Workload	Vessels Checked into ATAC	Remaining Workload	Response Rate
602 - Military Ships	267	0	267	127	140	47.57%
900 - Maritime/Merchant Vessels	1,153	8	1,159	570	589	49.18%
Total	**1,420**	**8**	**1,428**	**696**	**732**	**48.72%**

Shape your future START HERE >

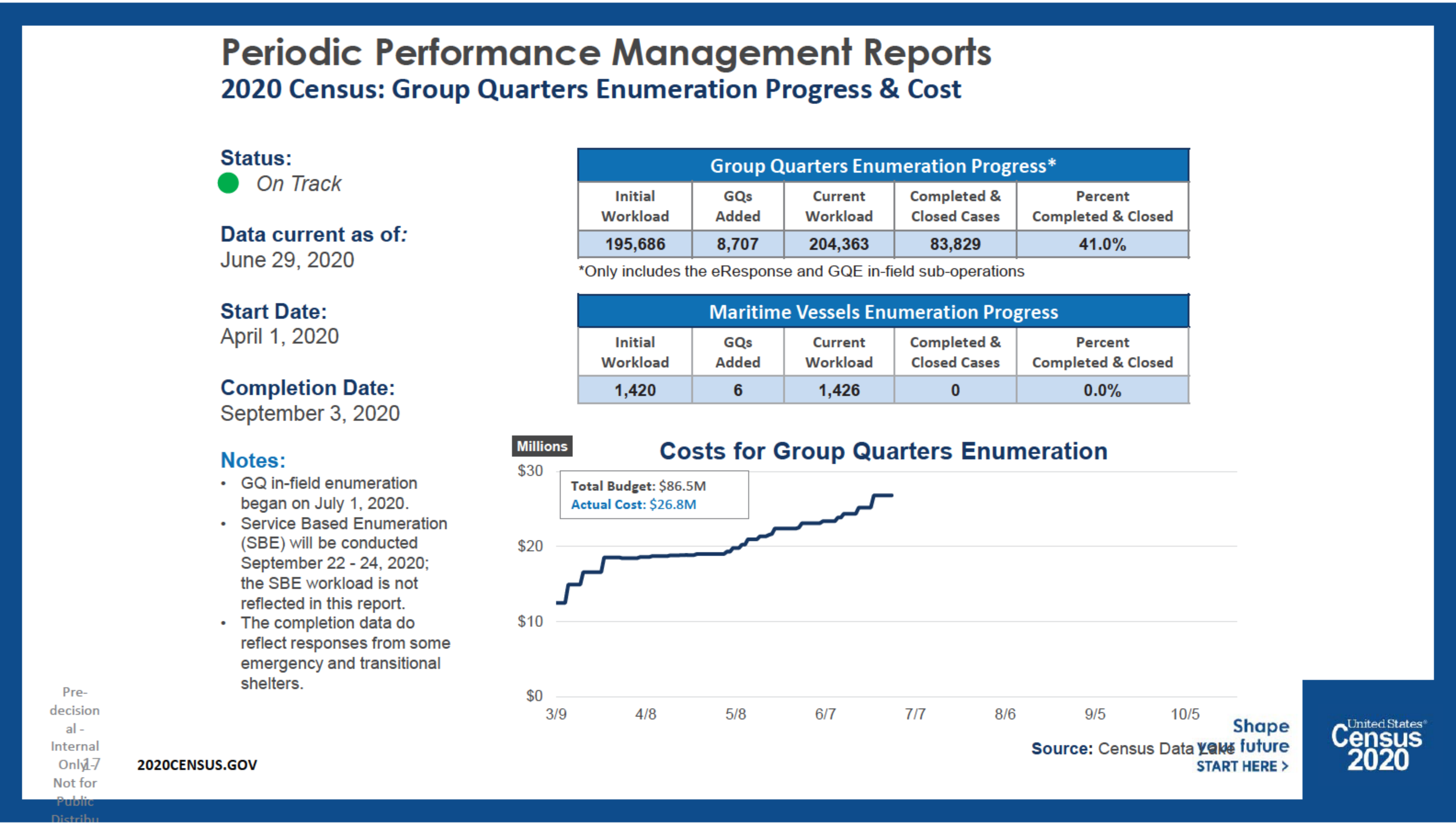
DRB Approval Number: CBDRB-FY21-DSEP-002



Backup

Schedules and Systems

DRB Approval Number: CBDRB-FY21-DSEP-002



Periodic Performance Management Reports

2020 Census: Group Quarters Enumeration Progress & Cost

Status:
● *On Track*

Data current as of:
June 29, 2020

Start Date:
April 1, 2020

Completion Date:
September 3, 2020

Notes:

- GQ in-field enumeration began on July 1, 2020.
- Service Based Enumeration (SBE) will be conducted September 22 - 24, 2020; the SBE workload is not reflected in this report.
- The completion data do reflect responses from some emergency and transitional shelters.

Group Quarters Enumeration Progress*				
Initial Workload	GQs Added	Current Workload	Completed & Closed Cases	Percent Completed & Closed
195,686	8,707	204,363	83,829	41.0%

*Only includes the eResponse and GQE in-field sub-operations

Maritime Vessels Enumeration Progress				
Initial Workload	GQs Added	Current Workload	Completed & Closed Cases	Percent Completed & Closed
1,420	6	1,426	0	0.0%

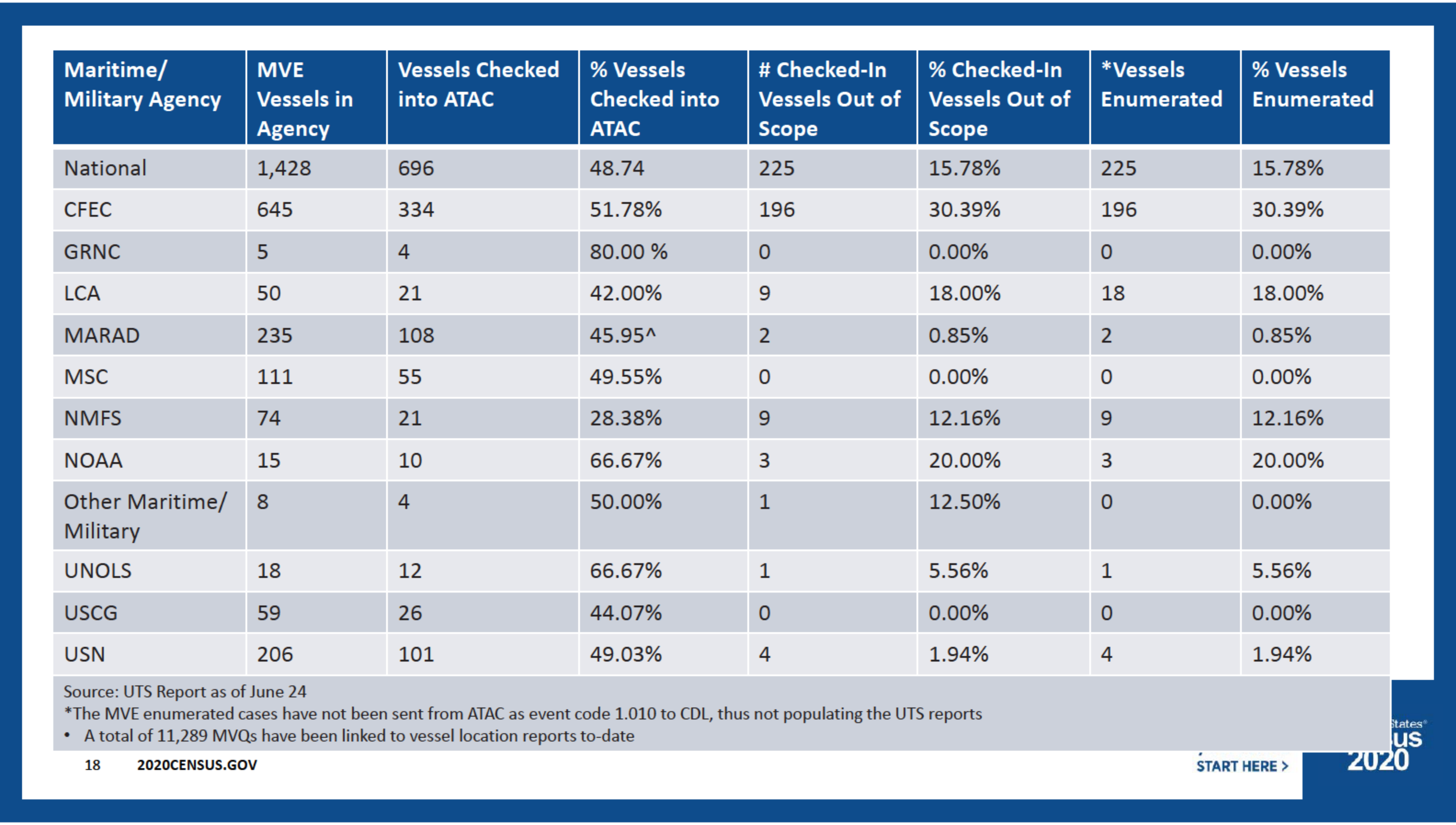
Costs for Group Quarters Enumeration

Total Budget: $86.5M
Actual Cost: $26.8M

Source: Census Data Lake

Pre-decisional - Internal Only Not for Public Distribu

DRB Approval Number: CBDRB-FY21-DSEP-002



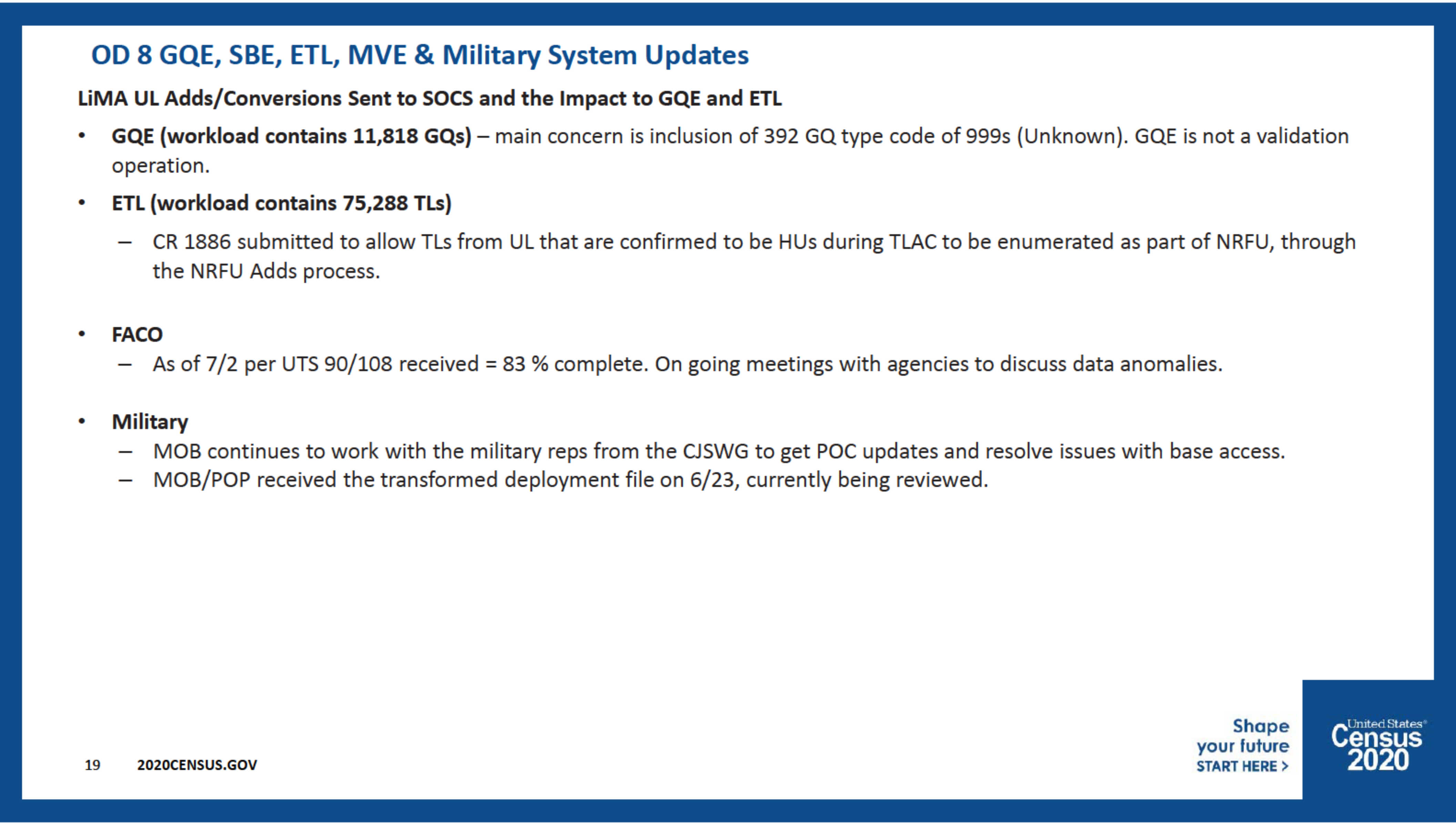
Maritime/ Military Agency	MVE Vessels in Agency	Vessels Checked into ATAC	% Vessels Checked into ATAC	# Checked-In Vessels Out of Scope	% Checked-In Vessels Out of Scope	*Vessels Enumerated	% Vessels Enumerated
National	1,428	696	48.74	225	15.78%	225	15.78%
CFEC	645	334	51.78%	196	30.39%	196	30.39%
GRNC	5	4	80.00 %	0	0.00%	0	0.00%
LCA	50	21	42.00%	9	18.00%	18	18.00%
MARAD	235	108	45.95^	2	0.85%	2	0.85%
MSC	111	55	49.55%	0	0.00%	0	0.00%
NMFS	74	21	28.38%	9	12.16%	9	12.16%
NOAA	15	10	66.67%	3	20.00%	3	20.00%
Other Maritime/ Military	8	4	50.00%	1	12.50%	0	0.00%
UNOLS	18	12	66.67%	1	5.56%	1	5.56%
USCG	59	26	44.07%	0	0.00%	0	0.00%
USN	206	101	49.03%	4	1.94%	4	1.94%

Source: UTS Report as of June 24

*The MVE enumerated cases have not been sent from ATAC as event code 1.010 to CDL, thus not populating the UTS reports

- A total of 11,289 MVQs have been linked to vessel location reports to-date

START HERE >



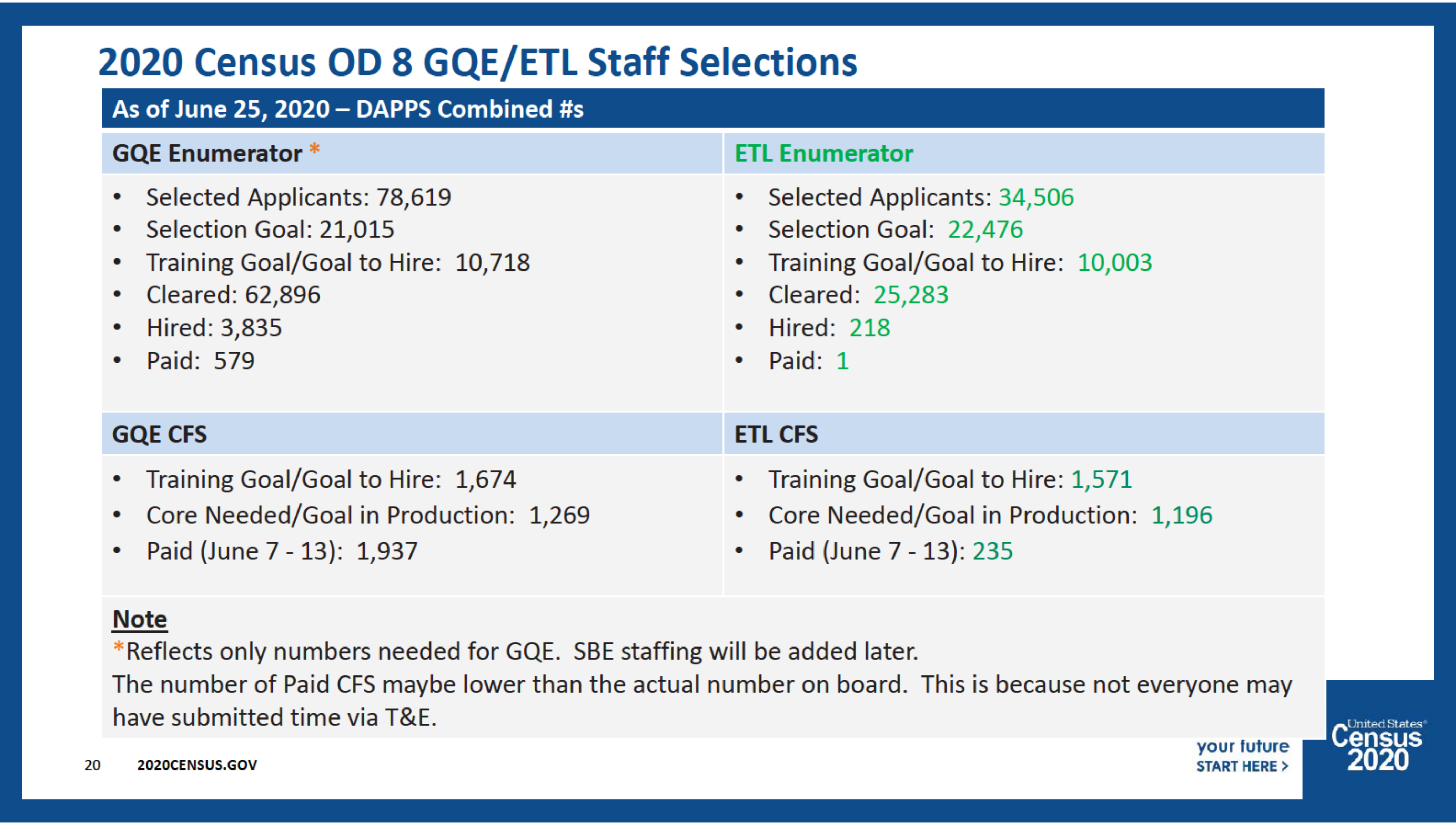
OD 8 GQE, SBE, ETL, MVE & Military System Updates

LiMA UL Adds/Conversions Sent to SOCS and the Impact to GQE and ETL

- **GQE (workload contains 11,818 GQs)** – main concern is inclusion of 392 GQ type code of 999s (Unknown). GQE is not a validation operation.
- **ETL (workload contains 75,288 TLs)**
  - CR 1886 submitted to allow TLs from UL that are confirmed to be HUs during TLAC to be enumerated as part of NRFU, through the NRFU Adds process.
- **FACO**
  - As of 7/2 per UTS 90/108 received = 83 % complete. On going meetings with agencies to discuss data anomalies.
- **Military**
  - MOB continues to work with the military reps from the CJSWG to get POC updates and resolve issues with base access.
  - MOB/POP received the transformed deployment file on 6/23, currently being reviewed.

Shape your future START HERE >

United States Census 2020



2020 Census OD 8 GQE/ETL Staff Selections

As of June 25, 2020 – DAPPS Combined #s

GQE Enumerator *

- Selected Applicants: 78,619
- Selection Goal: 21,015
- Training Goal/Goal to Hire: 10,718
- Cleared: 62,896
- Hired: 3,835
- Paid: 579

ETL Enumerator

- Selected Applicants: 34,506
- Selection Goal: 22,476
- Training Goal/Goal to Hire: 10,003
- Cleared: 25,283
- Hired: 218
- Paid: 1

GQE CFS

- Training Goal/Goal to Hire: 1,674
- Core Needed/Goal in Production: 1,269
- Paid (June 7 - 13): 1,937

ETL CFS

- Training Goal/Goal to Hire: 1,571
- Core Needed/Goal in Production: 1,196
- Paid (June 7 - 13): 235

Note

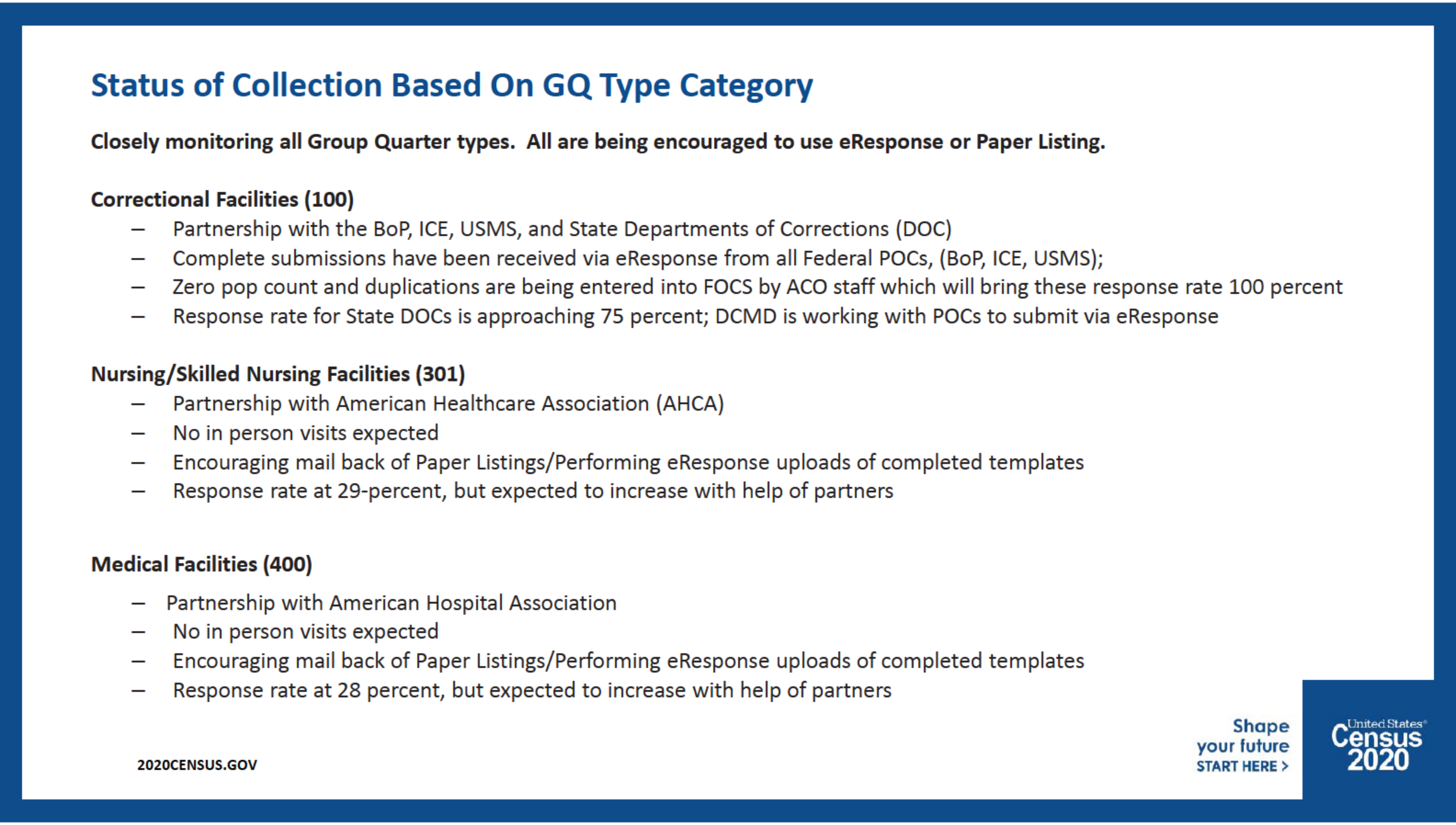
*Reflects only numbers needed for GQE. SBE staffing will be added later.

The number of Paid CFS maybe lower than the actual number on board. This is because not everyone may have submitted time via T&E.

your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



Status of Collection Based On GQ Type Category

Closely monitoring all Group Quarter types. All are being encouraged to use eResponse or Paper Listing.

Correctional Facilities (100)

- Partnership with the BoP, ICE, USMS, and State Departments of Corrections (DOC)
- Complete submissions have been received via eResponse from all Federal POCs, (BoP, ICE, USMS);
- Zero pop count and duplications are being entered into FOCS by ACO staff which will bring these response rate 100 percent
- Response rate for State DOCs is approaching 75 percent; DCMD is working with POCs to submit via eResponse

Nursing/Skilled Nursing Facilities (301)

- Partnership with American Healthcare Association (AHCA)
- No in person visits expected
- Encouraging mail back of Paper Listings/Performing eResponse uploads of completed templates
- Response rate at 29-percent, but expected to increase with help of partners

Medical Facilities (400)

- Partnership with American Hospital Association
- No in person visits expected
- Encouraging mail back of Paper Listings/Performing eResponse uploads of completed templates
- Response rate at 28 percent, but expected to increase with help of partners

2020CENSUS.GOV

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

Status of Collection Based On GQ Type Category

College/University Student Housing (501/502)

- Partnership with AACRAO, Department of Education
- Data for student housing steadily coming in
- Response rate close to 40 percent
- Response rate expected to be higher than data reveals due to administrators providing data for all of their students using only one Census ID/ User ID.
- Working with DSSD to determine the frequency of this action to obtain the data and work with ACOs/DCMD to split out the data across appropriate Census IDs

Military Barracks (601)

- Partnership with the Census Joint Services Working Group
- Received eResponse submissions for some barracks (Military Disciplinary Barracks)
- Military liaisons waiting for July to begin facility self enumeration, as planned earlier
- Encouraging eResponse or Paper Listing

Maritime/Military Vessels

- Partnership with the Vessel Working Group
- Working group continues to reach out to non-responding vessel operators to request return of completed Census data
- Mailed vessel reminders letter to non-responding vessel operators
- Response rate 48.88 percent

Status of Collection Based On GQ Type Category

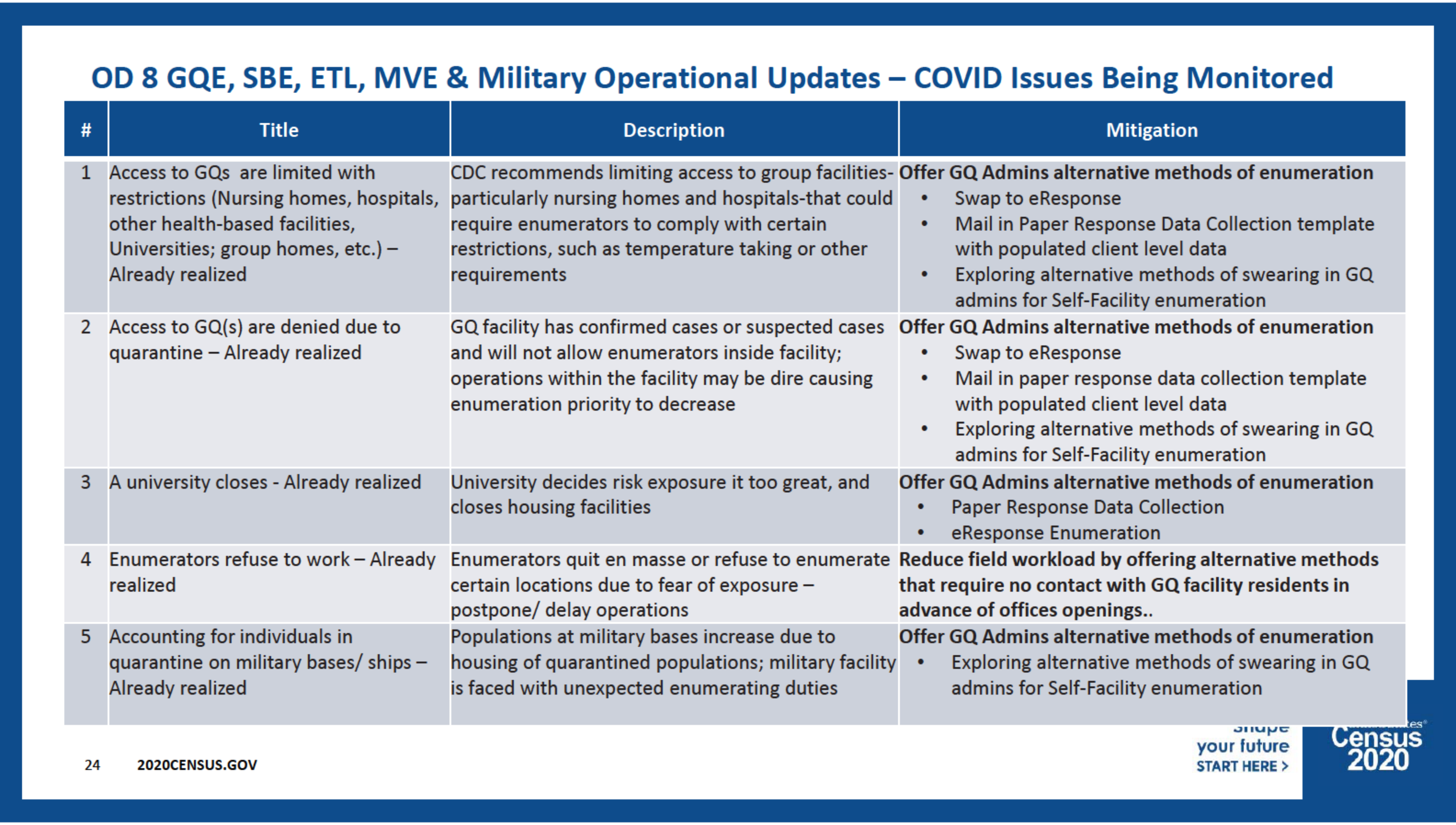
Group Homes/Residential Treatment Centers (Juvenile/Adults) (200/800)

- Worked with Legal to obtain a standard HIPAA compliance letter to share with juvenile and treatment facilities concerned about releasing health care information
- Includes language received from the Department of Health and Human Services regarding the Privacy Rule, under the Health Insurance Portability and Accountability Act of 1996 (HIPPA), which addresses the use and disclosure of individual health information. The Privacy Rule also sets standards for individuals' privacy rights, to understand and control how their health information is used. Since the decennial census does not request any health information, neither HIPPA nor the Privacy Rule bars response.
- Assisting GQ administrators with data uploads to minimize level of effort and stress during this pandemic

Service-Based Facilities (700)

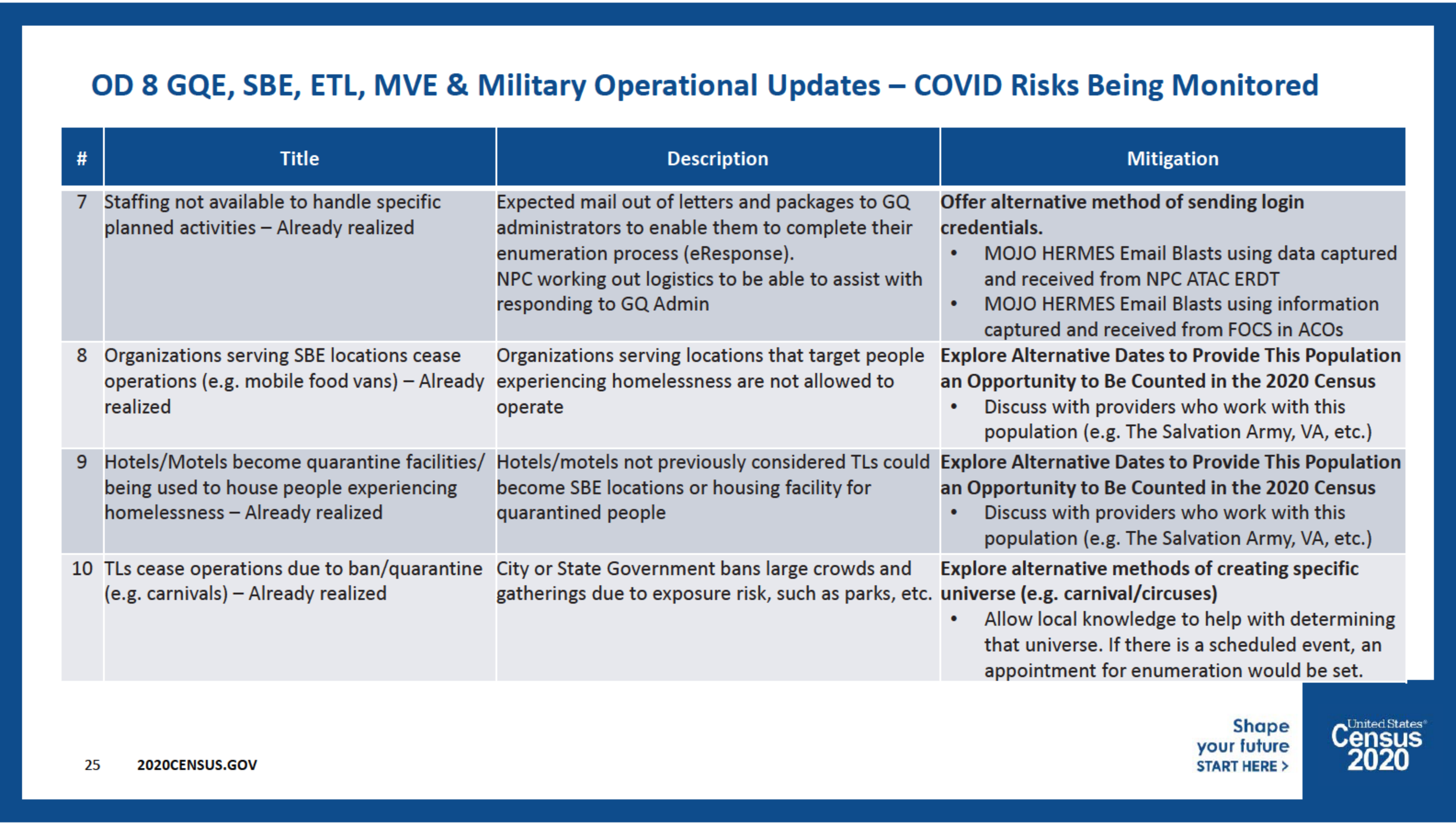
- Three day In-person enumeration to began 9/22

All GQ types with the exception of SBEs are being encouraged to use eResponse or Paper Listing and closely monitored.



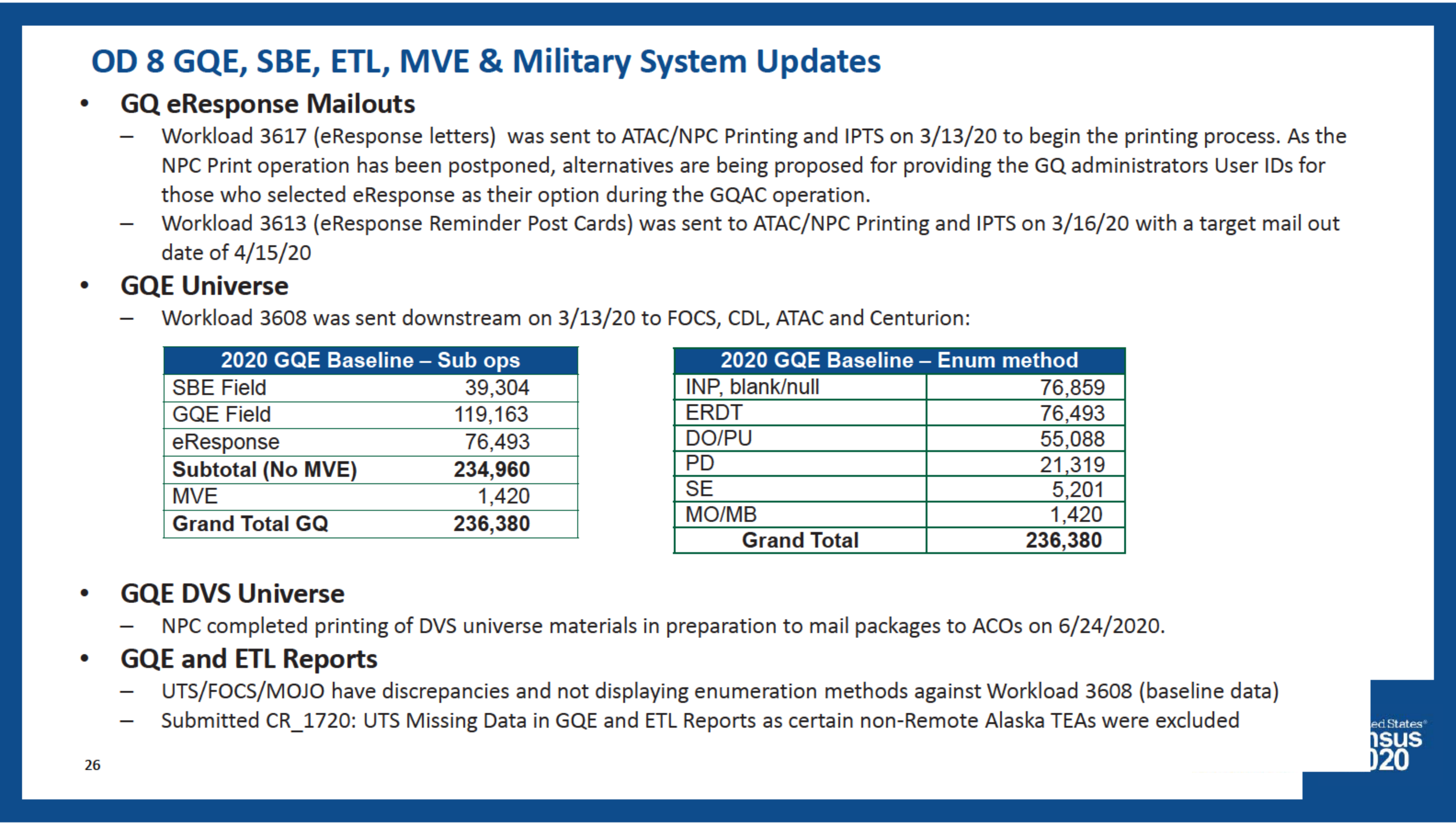
OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Issues Being Monitored

#	Title	Description	Mitigation
1	Access to GQs are limited with restrictions (Nursing homes, hospitals, other health-based facilities, Universities; group homes, etc.) – Already realized	CDC recommends limiting access to group facilities-particularly nursing homes and hospitals-that could require enumerators to comply with certain restrictions, such as temperature taking or other requirements	**Offer GQ Admins alternative methods of enumeration** • Swap to eResponse • Mail in Paper Response Data Collection template with populated client level data • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration
2	Access to GQ(s) are denied due to quarantine – Already realized	GQ facility has confirmed cases or suspected cases and will not allow enumerators inside facility; operations within the facility may be dire causing enumeration priority to decrease	**Offer GQ Admins alternative methods of enumeration** • Swap to eResponse • Mail in paper response data collection template with populated client level data • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration
3	A university closes - Already realized	University decides risk exposure it too great, and closes housing facilities	**Offer GQ Admins alternative methods of enumeration** • Paper Response Data Collection • eResponse Enumeration
4	Enumerators refuse to work – Already realized	Enumerators quit en masse or refuse to enumerate certain locations due to fear of exposure – postpone/ delay operations	**Reduce field workload by offering alternative methods that require no contact with GQ facility residents in advance of offices openings..**
5	Accounting for individuals in quarantine on military bases/ ships – Already realized	Populations at military bases increase due to housing of quarantined populations; military facility is faced with unexpected enumerating duties	**Offer GQ Admins alternative methods of enumeration** • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration



OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Risks Being Monitored

#	Title	Description	Mitigation
7	Staffing not available to handle specific planned activities – Already realized	Expected mail out of letters and packages to GQ administrators to enable them to complete their enumeration process (eResponse). NPC working out logistics to be able to assist with responding to GQ Admin	**Offer alternative method of sending login credentials.** • MOJO HERMES Email Blasts using data captured and received from NPC ATAC ERDT • MOJO HERMES Email Blasts using information captured and received from FOCS in ACOs
8	Organizations serving SBE locations cease operations (e.g. mobile food vans) – Already realized	Organizations serving locations that target people experiencing homelessness are not allowed to operate	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
9	Hotels/Motels become quarantine facilities/ being used to house people experiencing homelessness – Already realized	Hotels/motels not previously considered TLs could become SBE locations or housing facility for quarantined people	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
10	TLs cease operations due to ban/quarantine (e.g. carnivals) – Already realized	City or State Government bans large crowds and gatherings due to exposure risk, such as parks, etc.	**Explore alternative methods of creating specific universe (e.g. carnival/circuses)** • Allow local knowledge to help with determining that universe. If there is a scheduled event, an appointment for enumeration would be set.



OD 8 GQE, SBE, ETL, MVE & Military System Updates

- **GQ eResponse Mailouts**
  - Workload 3617 (eResponse letters) was sent to ATAC/NPC Printing and IPTS on 3/13/20 to begin the printing process. As the NPC Print operation has been postponed, alternatives are being proposed for providing the GQ administrators User IDs for those who selected eResponse as their option during the GQAC operation.
  - Workload 3613 (eResponse Reminder Post Cards) was sent to ATAC/NPC Printing and IPTS on 3/16/20 with a target mail out date of 4/15/20
- **GQE Universe**
  - Workload 3608 was sent downstream on 3/13/20 to FOCS, CDL, ATAC and Centurion:

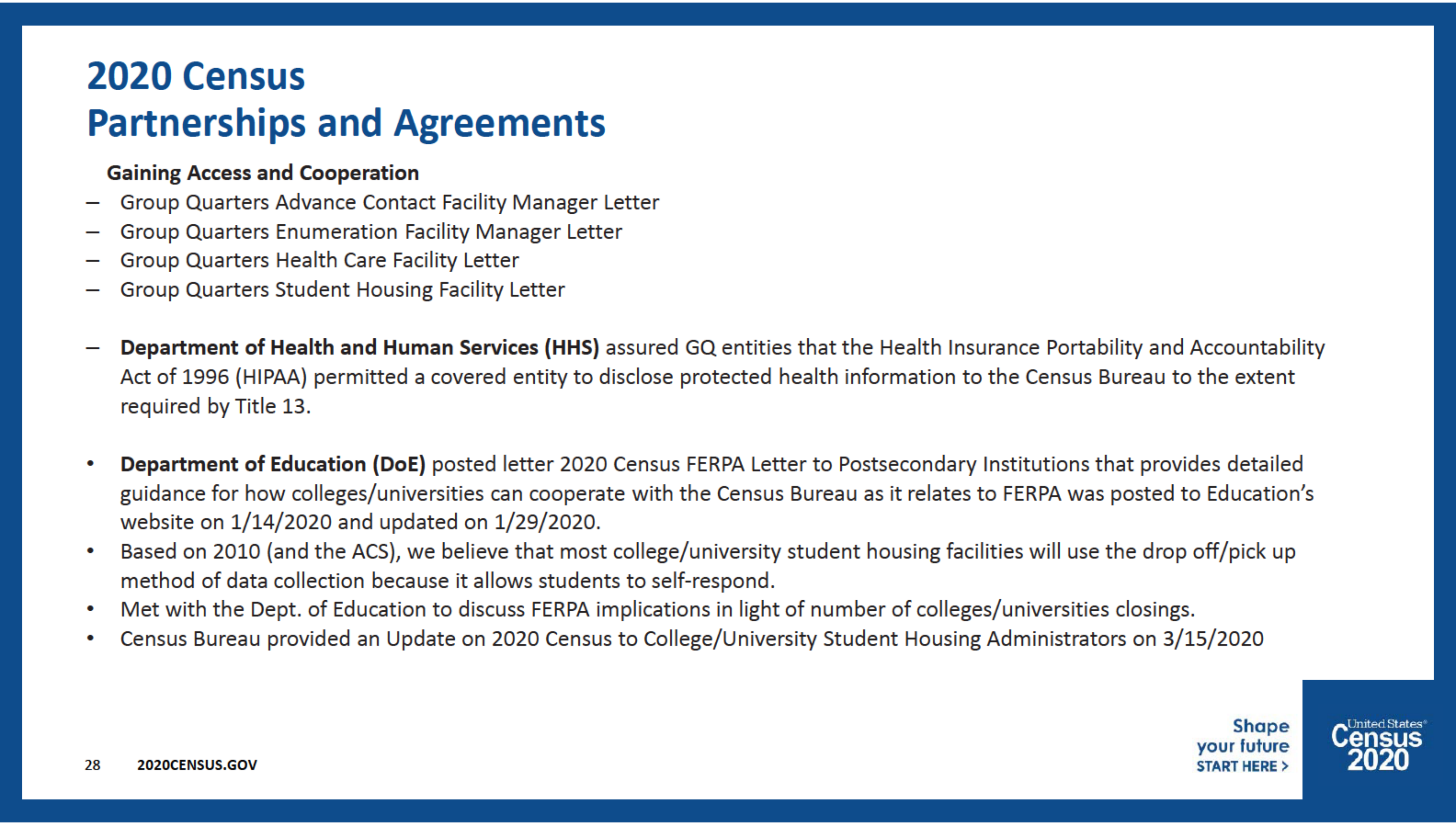
2020 GQE Baseline – Sub ops	
SBE Field	39,304
GQE Field	119,163
eResponse	76,493
Subtotal (No MVE)	**234,960**
MVE	1,420
Grand Total GQ	**236,380**

2020 GQE Baseline – Enum method	
INP, blank/null	76,859
ERDT	76,493
DO/PU	55,088
PD	21,319
SE	5,201
MO/MB	1,420
Grand Total	**236,380**

- **GQE DVS Universe**
  - NPC completed printing of DVS universe materials in preparation to mail packages to ACOs on 6/24/2020.
- **GQE and ETL Reports**
  - UTS/FOCS/MOJO have discrepancies and not displaying enumeration methods against Workload 3608 (baseline data)
  - Submitted CR_1720: UTS Missing Data in GQE and ETL Reports as certain non-Remote Alaska TEAs were excluded

OD 8 GQE, SBE, ETL, MVE & Military Operational Outreach Throughout COVID

- March – Nationwide “Stay at Home” orders
- **March:** Letter developed by DCMD for RCCs to send to GQs that selected self enumeration options to change option to eResponse or paper listings
- **March 13:** Posted letter, *Update on 2020 Census for Student Housing Administrators* on the Census Bureau and Department of Education Website, requesting administrators who selected self enumeration option to change option to eResponse or Paper Listings
- **March 25:** DCMD sent *Update on the 2020 Census for Health Care Administrator* letter to Health Care umbrella organizations providing guidance for Administrators that selected a self-enumeration option to change method of enumeration to eResponse of Paper Listings
- NPC/ Jeffersonville call center was closed due to COVID and was not available to complete task in support of eResponse. As a result:
  - NPC was unable to meet the March 27 deadline for mailing eResponse Letters with Login credentials
  - DCMD worked with NPC ATAC management to update the system to allow multiple users to view and update email address
  - DCMD stood up a Call Center with staff across ACOs and Census HQ volunteers to verify/update email address to deliver login credentials.
- **March 31:** Mojo/Hermes sent out 1st email blast with login credentials to GQ admins who has selected eResponse during GQAC
- **April 1:** GQE eResponse portal became available for GQ submittals.
- **April 13** and **April 20:** Mojo/Hermes sent out email blast 2 and 3 with login credentials for bounce back emails from 1st email blast
- **April 2:** DCMD Staff and volunteers across the Decennial Directorate and other started reaching out in response to questions from GQ administrators received via email and phone calls.
- April ? PIO developed a video to college students. Posted on website encouraging internet response or that GQ admins would respond for them if they live in student housing
- **April 20:** ACO began calling GQ admins to offer Mail out/ Mail back Paper Response Data Collection (Paper Listings)
- **May 28:** Census Bureau participated in a webinar to remind/ update student housing administrators on the 2020 Census Group Quarters operation and to inform administrators of the upcoming request for off-campus student data.
- **June 3:** NPC mailed Maritime/Military Vessel reminder letters to non-responding vessel operators
- **June 8:** ACO staff began calling GQ administrators to reschedule appointment dates for their facilities.
- **June 11:** Meeting with AACRAO…
- **June 22:** Provided updates to The Salvation Army for their Directive to be sent to their managing entities.
- **June 22:** Met with National Network to End Domestic Violence to discuss upcoming enumeration, options, and COVID-19 procedures



2020 Census
Partnerships and Agreements

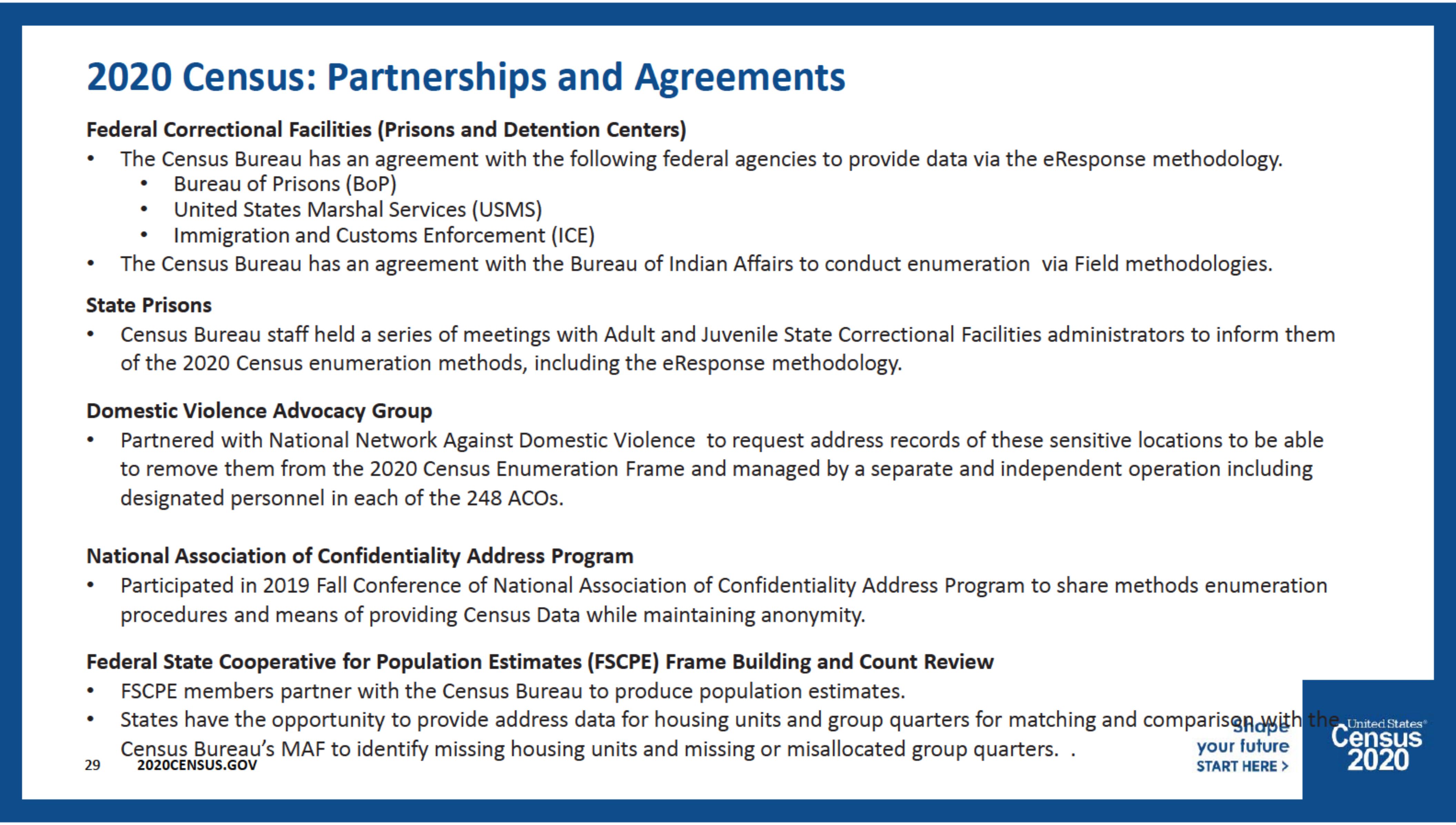
Gaining Access and Cooperation

- Group Quarters Advance Contact Facility Manager Letter
- Group Quarters Enumeration Facility Manager Letter
- Group Quarters Health Care Facility Letter
- Group Quarters Student Housing Facility Letter

- **Department of Health and Human Services (HHS)** assured GQ entities that the Health Insurance Portability and Accountability Act of 1996 (HIPAA) permitted a covered entity to disclose protected health information to the Census Bureau to the extent required by Title 13.

- **Department of Education (DoE)** posted letter 2020 Census FERPA Letter to Postsecondary Institutions that provides detailed guidance for how colleges/universities can cooperate with the Census Bureau as it relates to FERPA was posted to Education's website on 1/14/2020 and updated on 1/29/2020.
- Based on 2010 (and the ACS), we believe that most college/university student housing facilities will use the drop off/pick up method of data collection because it allows students to self-respond.
- Met with the Dept. of Education to discuss FERPA implications in light of number of colleges/universities closings.
- Census Bureau provided an Update on 2020 Census to College/University Student Housing Administrators on 3/15/2020

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census: Partnerships and Agreements

Federal Correctional Facilities (Prisons and Detention Centers)

- The Census Bureau has an agreement with the following federal agencies to provide data via the eResponse methodology.
  - Bureau of Prisons (BoP)
  - United States Marshal Services (USMS)
  - Immigration and Customs Enforcement (ICE)
- The Census Bureau has an agreement with the Bureau of Indian Affairs to conduct enumeration via Field methodologies.

State Prisons

- Census Bureau staff held a series of meetings with Adult and Juvenile State Correctional Facilities administrators to inform them of the 2020 Census enumeration methods, including the eResponse methodology.

Domestic Violence Advocacy Group

- Partnered with National Network Against Domestic Violence to request address records of these sensitive locations to be able to remove them from the 2020 Census Enumeration Frame and managed by a separate and independent operation including designated personnel in each of the 248 ACOs.

National Association of Confidentiality Address Program

- Participated in 2019 Fall Conference of National Association of Confidentiality Address Program to share methods enumeration procedures and means of providing Census Data while maintaining anonymity.

Federal State Cooperative for Population Estimates (FSCPE) Frame Building and Count Review

- FSCPE members partner with the Census Bureau to produce population estimates.
- States have the opportunity to provide address data for housing units and group quarters for matching and comparison with the Census Bureau's MAF to identify missing housing units and missing or misallocated group quarters. .

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census
Partnerships and Agreements

Salvation Army

- The Salvation completed and will provide each of their entities with the appropriate Directive.
  - 2020 Decennial Population Census – General (Handling of Soup kitchens and mobile food vans)
  - 2020 Decennial Population Census – Residential Institutions (Adult Rehabilitation Centers, Harbor Light Centers, Transient Lodges, residential facilities for children and other temporary housing facilities such as shelters for men, women and/or families)

American Healthcare Association (AHCA)

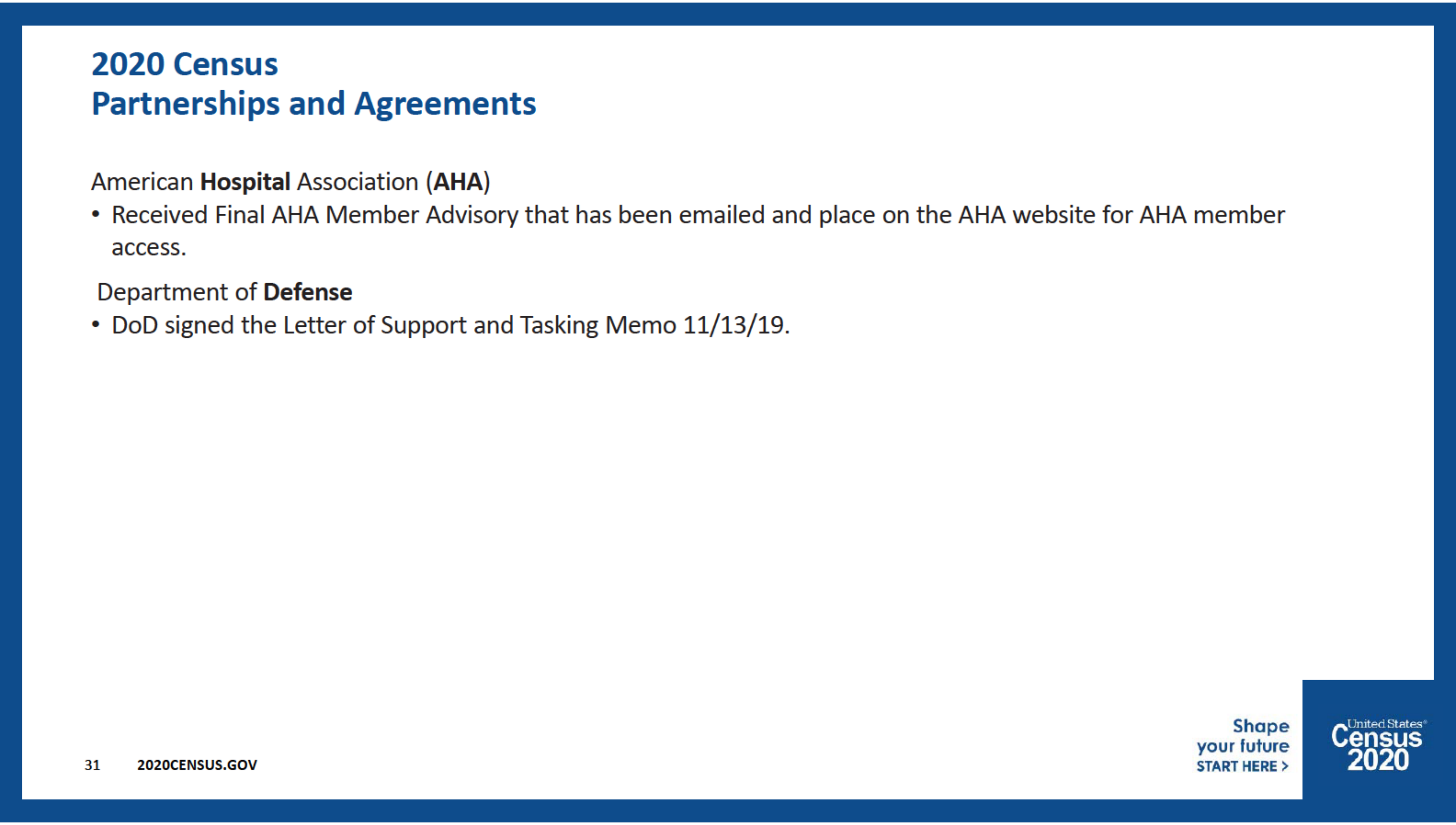
- The Census Bureau provided final updates for the 2020 Regulatory Advisory Letter to be sent to entities of the AHCA and the National Center for Assisted Living (NCAL) to inform them of 2020 Census enumeration plans.
- The Census Bureau provided additional feedback to the 2020 Census Procedures Outlined for Long Term Care (LTC) Facilities from AHCA who also had questions about classifying their Intermediate Care Facilities for Individuals with Intellectual Disabilities Facilities (ICFs/ID) as either GQs and housing units and concerns about the health of their residents and census workers in light of the COVID-19 Outbreak. POP provided language on what health care facilities could expect, depending on which operation they had already been assigned to.

Veteran Affairs

- The Veteran Affairs/ Administration provided Census with address records of location where they house veterans who may be experiencing homelessness.

Shape your future START HERE >

United States Census 2020



2020 Census
Partnerships and Agreements

American **Hospital** Association (**AHA**)

- Received Final AHA Member Advisory that has been emailed and place on the AHA website for AHA member access.

Department of **Defense**

- DoD signed the Letter of Support and Tasking Memo 11/13/19.

Shape your future START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002

OD 8 GQE, SBE, ETL, MVE & Military Operational Updates

2020census.gov website

Conducting the Count: https://2020census.gov/en/conducting-the-count.html

Counting People in Group Living Arrangements: https://2020census.gov/en/conducting-the-count/gq.html

Group Quarters Enumeration: https://2020census.gov/en/conducting-the-count/gq/gqe.html

Service-Based Enumeration: https://2020census.gov/en/conducting-the-count/gq/sbe.html

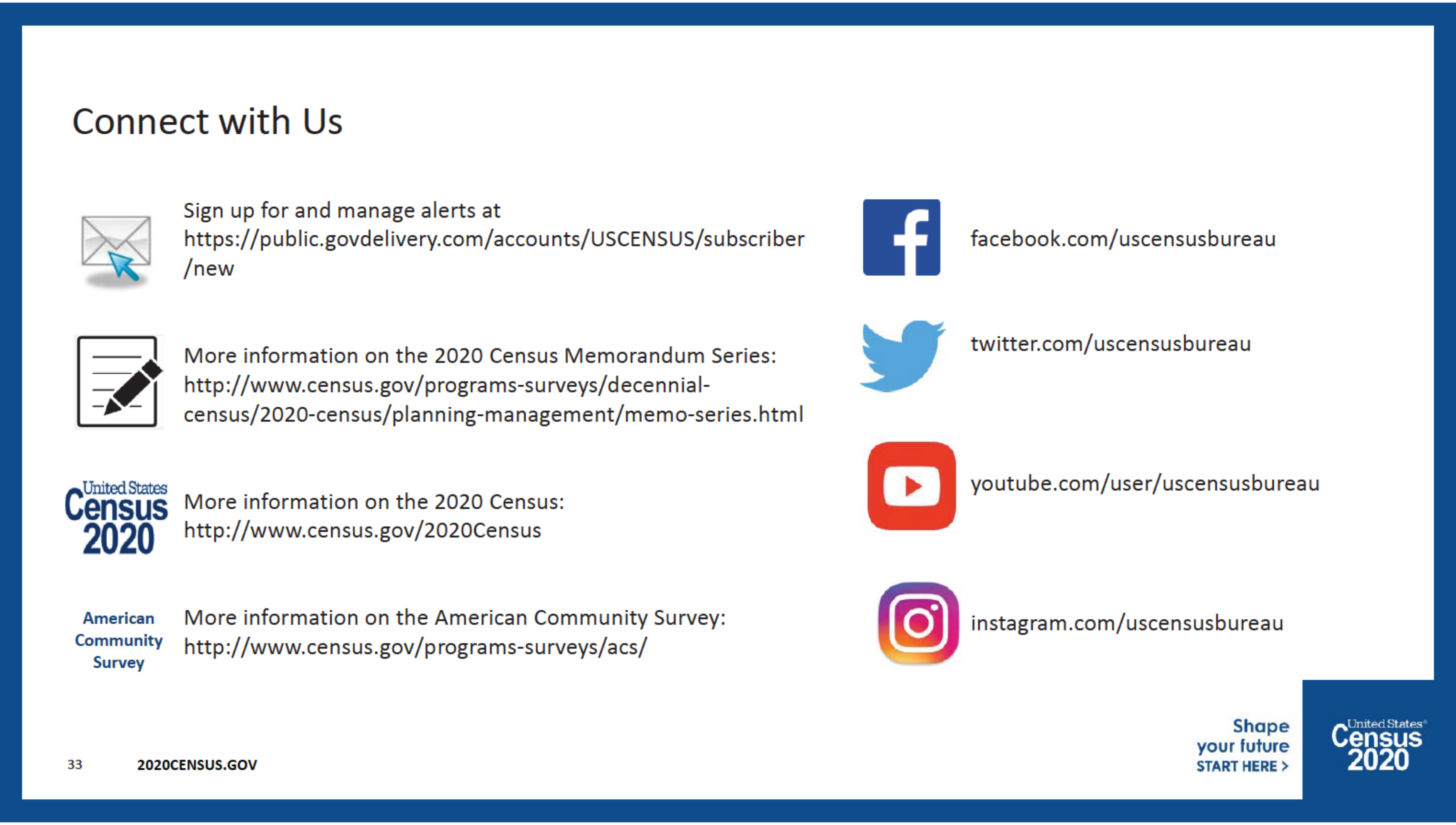
Group Quarters Advance Contact: https://2020census.gov/en/conducting-the-count/gq/gqac.html

eResponse: https://2020census.gov/en/conducting-the-count/gq/eresponse.html

Maritime and Military Vessel Enumeration: https://2020census.gov/en/conducting-the-count/gq/mve.html

Department of Education Student Privacy Policy Office: https://studentprivacy.ed.gov/faq/colleges-and-2020-census

DRB Approval Number: CBDRB-FY21-DSEP-002



Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

United States Census 2020

More information on the 2020 Census: http://www.census.gov/2020Census

American Community Survey

More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

instagram.com/uscensusbureau

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Operational Delivery 8: Group Quarters Update

Focus: Group Quarters Enumeration and Maritime/Military Vessel Enumeration

Thursday: July 2, 2020

Presented by: Dora Durante and Crystal Miller

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

Shape
your future
START HERE >

United States®
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Group Quarters Bottom Line Up Front (BLUF)

Update as of July 1, 2020

- **Group Quarters Workload excluding SBE and MVE: 204,433, Current Workload: 117,576**
- **Overall GQ level response rate is based on submissions that have been "checked in" via FOCS and ATAC. The response rate will be higher or lower based on group quarters type.**
- **Starting April 1 through June 30, we received data mainly from eResponse and Paper Listings**
- **GQ In-Person enumeration began July 1 and will continue through September 3, 2020**

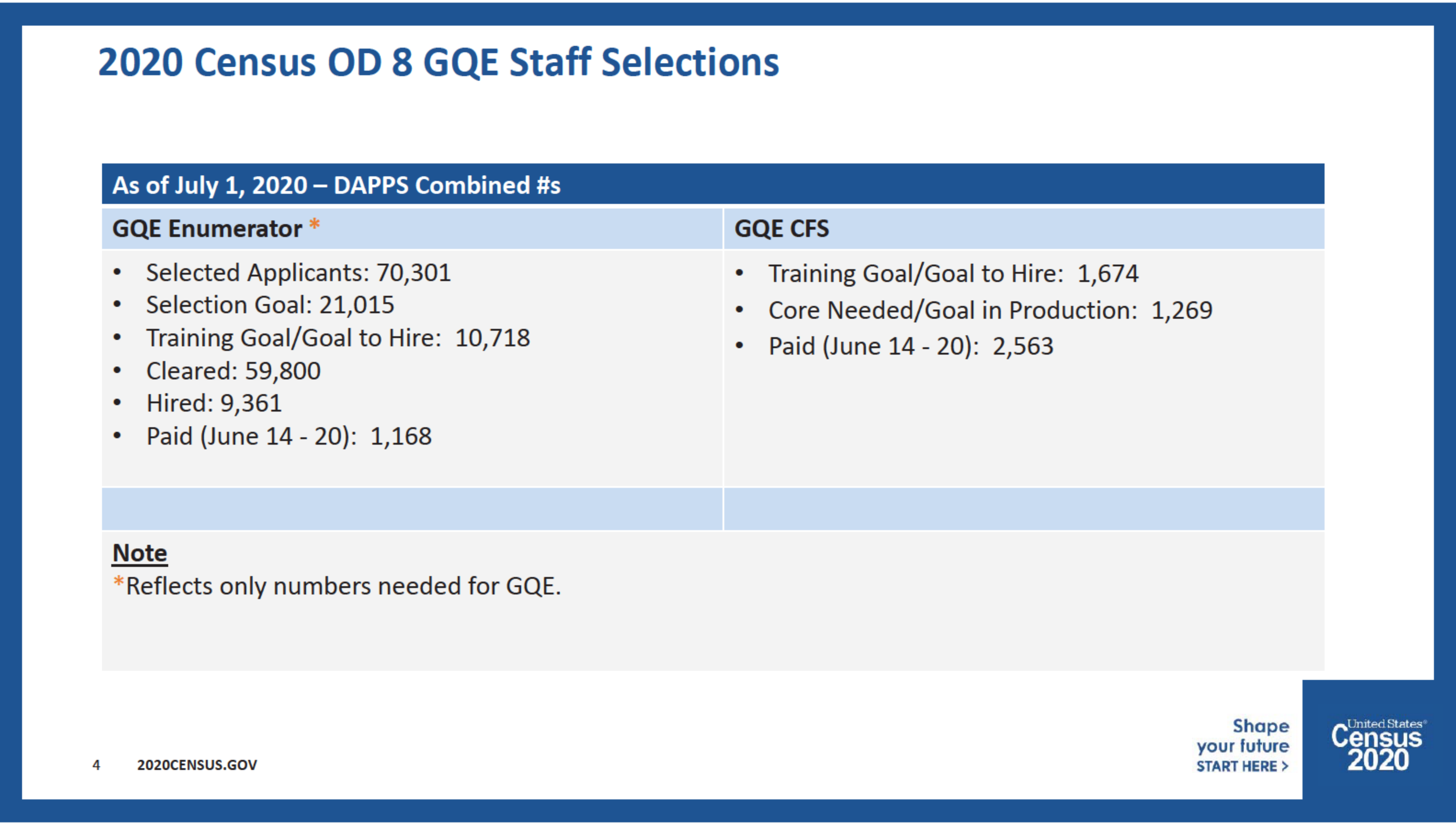
DRB Approval Number: CBDRB-FY21-DSEP-002

OD 8 GQE Milestone Conduct & Training Dates

Conduct Activity	Proposed Start	Proposed Finish
Conduct GQE eResponse Operation	**04/01/20 (A)**	**08/07/20 (P)**
Conduct GQE Operation Field	**04/20/20 (A)**	**08/26/20 (P)**
Conduct GQE In-Person Operation	**07/01/20 (A)**	08/26/20 (P)
Conduct GQ Reinterview Operation	07/02/20 (P)	09/03/20 (P)

Training Activity	Proposed Start	Proposed Finish
Conduct GQE CFS Refresher Training	**06/16/20 (A)**	**06/19/20 (A)**
Conduct GQE Clerk Refresher Training	**06/19/20 (A)**	
Conduct GQE Enumerator Training	**06/26/20 (A)**	**07/01/20 (A)**

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census OD 8 GQE Staff Selections

As of July 1, 2020 – DAPPS Combined #s

GQE Enumerator *	GQE CFS
• Selected Applicants: 70,301 • Selection Goal: 21,015 • Training Goal/Goal to Hire: 10,718 • Cleared: 59,800 • Hired: 9,361 • Paid (June 14 - 20): 1,168	• Training Goal/Goal to Hire: 1,674 • Core Needed/Goal in Production: 1,269 • Paid (June 14 - 20): 2,563

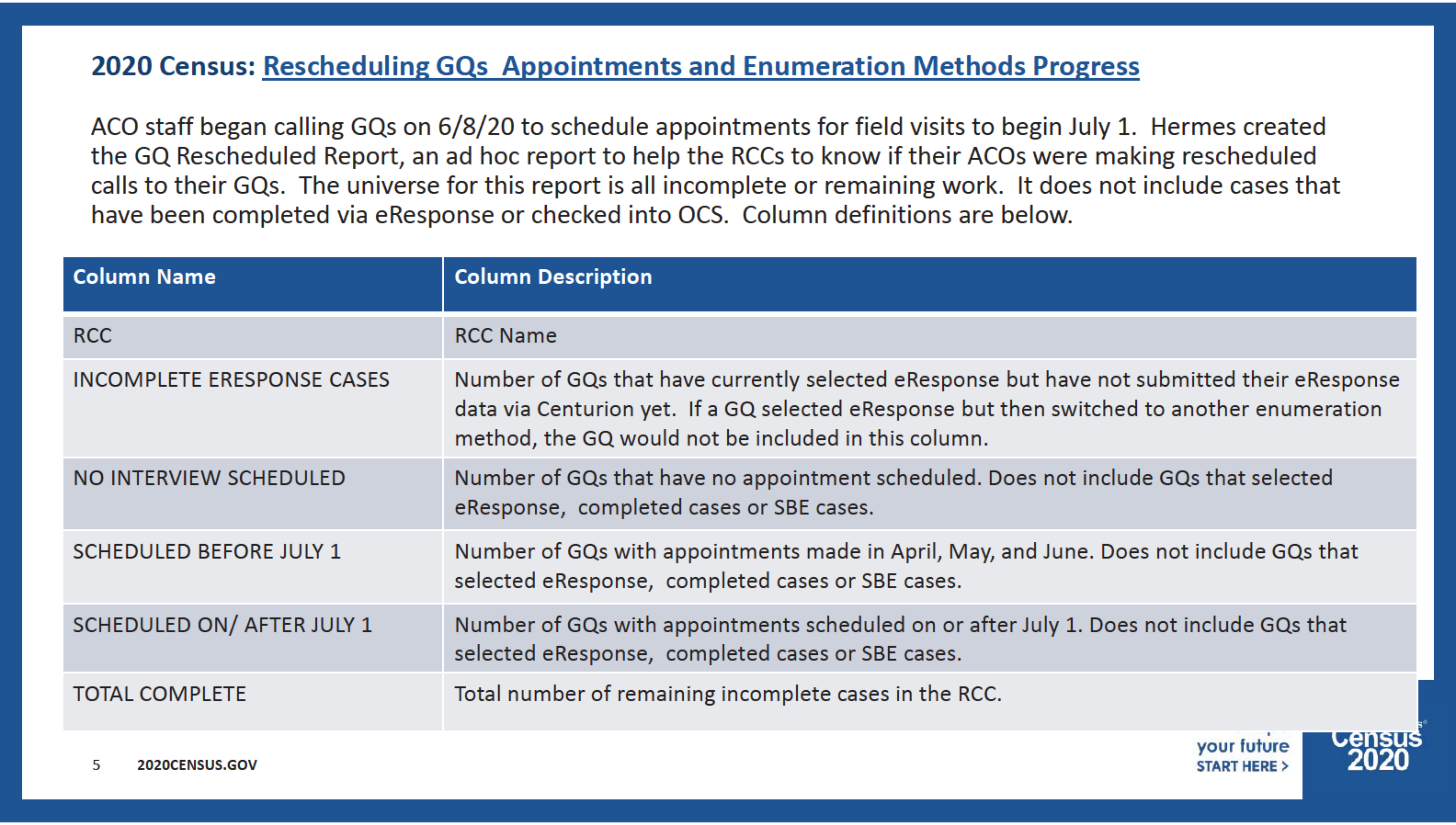
Note

*Reflects only numbers needed for GQE.

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

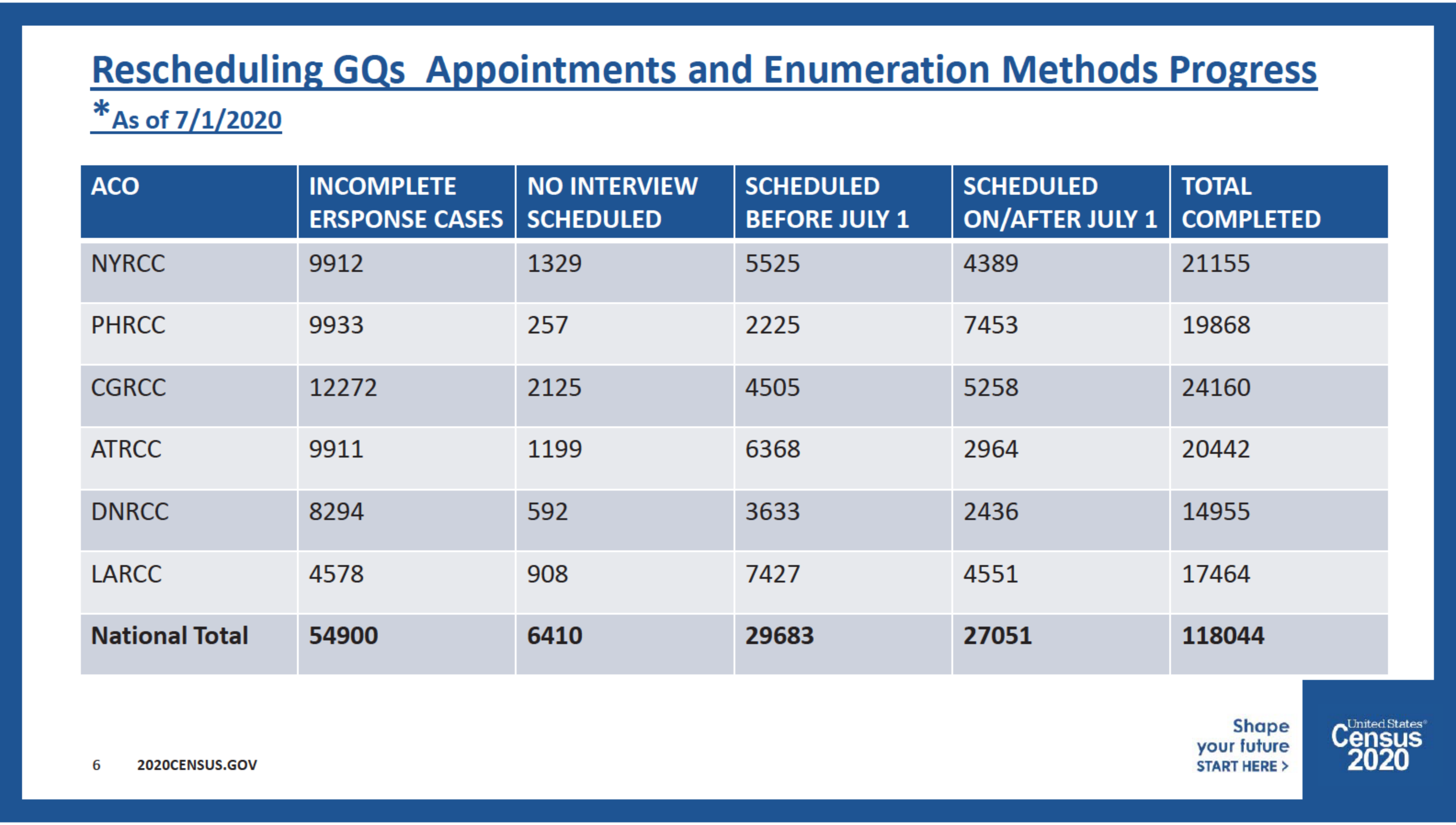


2020 Census: Rescheduling GQs Appointments and Enumeration Methods Progress

ACO staff began calling GQs on 6/8/20 to schedule appointments for field visits to begin July 1. Hermes created the GQ Rescheduled Report, an ad hoc report to help the RCCs to know if their ACOs were making rescheduled calls to their GQs. The universe for this report is all incomplete or remaining work. It does not include cases that have been completed via eResponse or checked into OCS. Column definitions are below.

Column Name	Column Description
RCC	RCC Name
INCOMPLETE ERESPONSE CASES	Number of GQs that have currently selected eResponse but have not submitted their eResponse data via Centurion yet. If a GQ selected eResponse but then switched to another enumeration method, the GQ would not be included in this column.
NO INTERVIEW SCHEDULED	Number of GQs that have no appointment scheduled. Does not include GQs that selected eResponse, completed cases or SBE cases.
SCHEDULED BEFORE JULY 1	Number of GQs with appointments made in April, May, and June. Does not include GQs that selected eResponse, completed cases or SBE cases.
SCHEDULED ON/ AFTER JULY 1	Number of GQs with appointments scheduled on or after July 1. Does not include GQs that selected eResponse, completed cases or SBE cases.
TOTAL COMPLETE	Total number of remaining incomplete cases in the RCC.

DRB Approval Number: CBDRB-FY21-DSEP-002

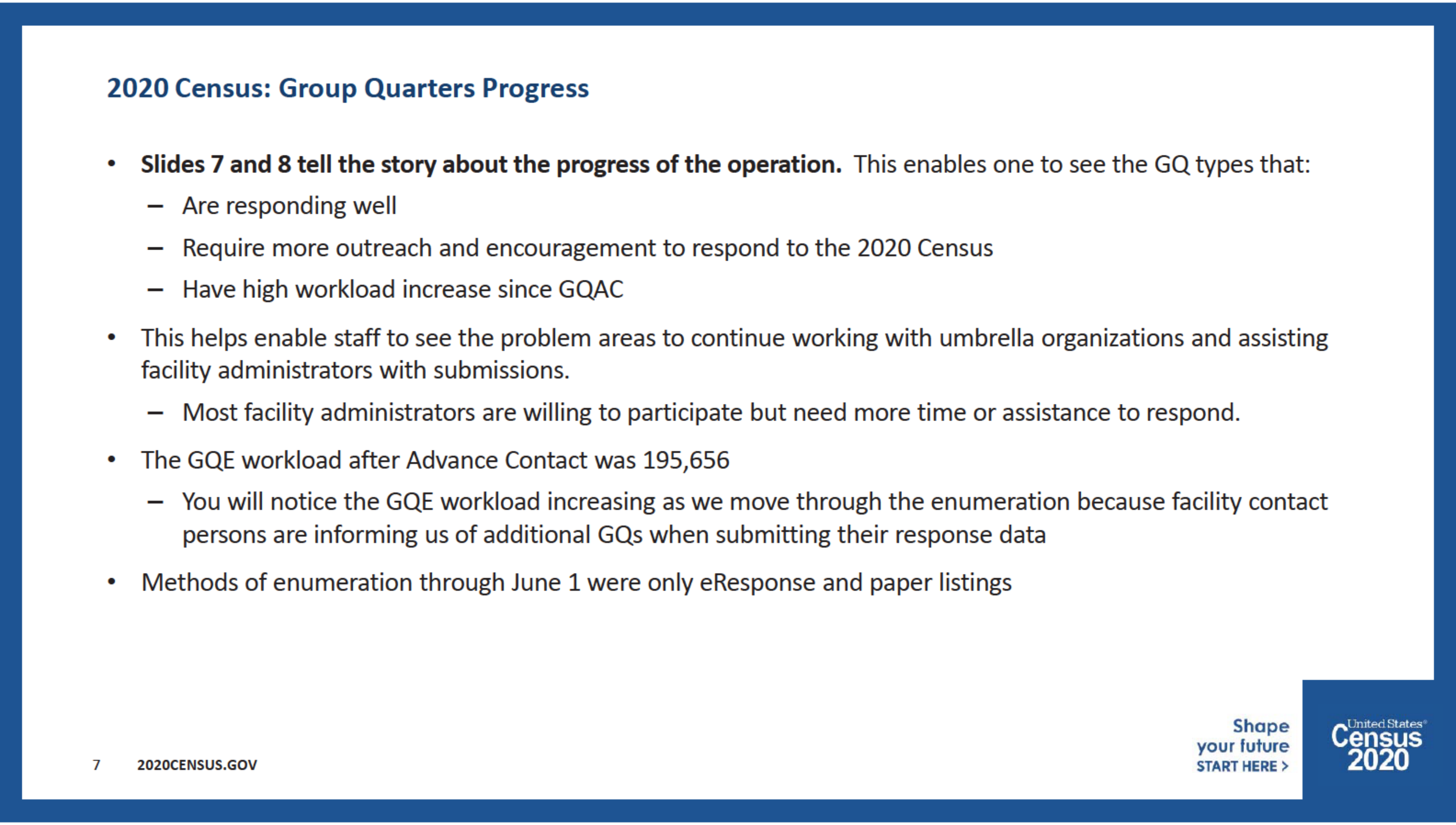


Rescheduling GQs Appointments and Enumeration Methods Progress

*As of 7/1/2020

ACO	INCOMPLETE ERSPONSE CASES	NO INTERVIEW SCHEDULED	SCHEDULED BEFORE JULY 1	SCHEDULED ON/AFTER JULY 1	TOTAL COMPLETED
NYRCC	9912	1329	5525	4389	21155
PHRCC	9933	257	2225	7453	19868
CGRCC	12272	2125	4505	5258	24160
ATRCC	9911	1199	6368	2964	20442
DNRCC	8294	592	3633	2436	14955
LARCC	4578	908	7427	4551	17464
National Total	**54900**	**6410**	**29683**	**27051**	**118044**

Shape your future
START HERE >



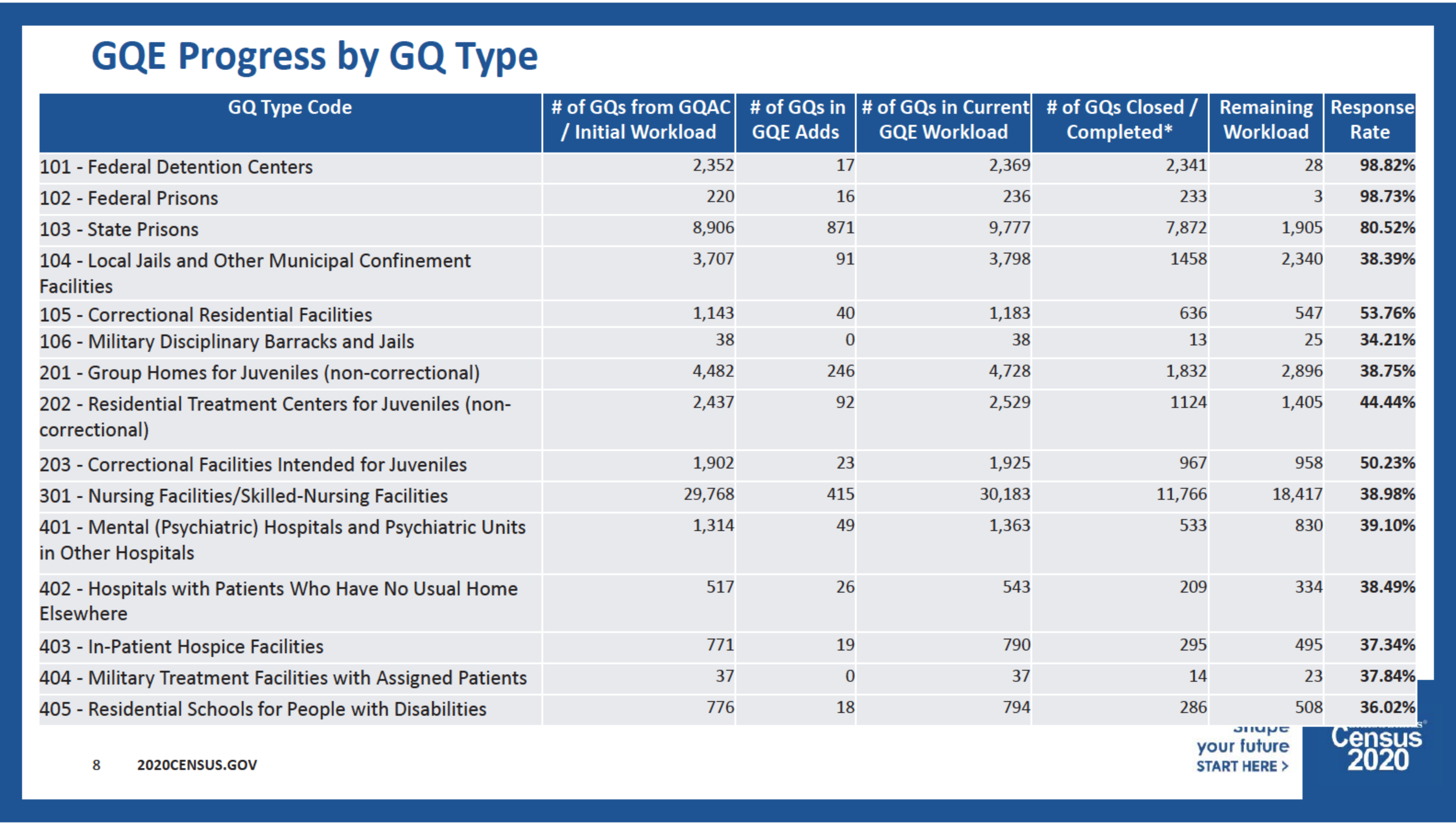
2020 Census: Group Quarters Progress

- **Slides 7 and 8 tell the story about the progress of the operation.** This enables one to see the GQ types that:
  - Are responding well
  - Require more outreach and encouragement to respond to the 2020 Census
  - Have high workload increase since GQAC
- This helps enable staff to see the problem areas to continue working with umbrella organizations and assisting facility administrators with submissions.
  - Most facility administrators are willing to participate but need more time or assistance to respond.
- The GQE workload after Advance Contact was 195,656
  - You will notice the GQE workload increasing as we move through the enumeration because facility contact persons are informing us of additional GQs when submitting their response data
- Methods of enumeration through June 1 were only eResponse and paper listings

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



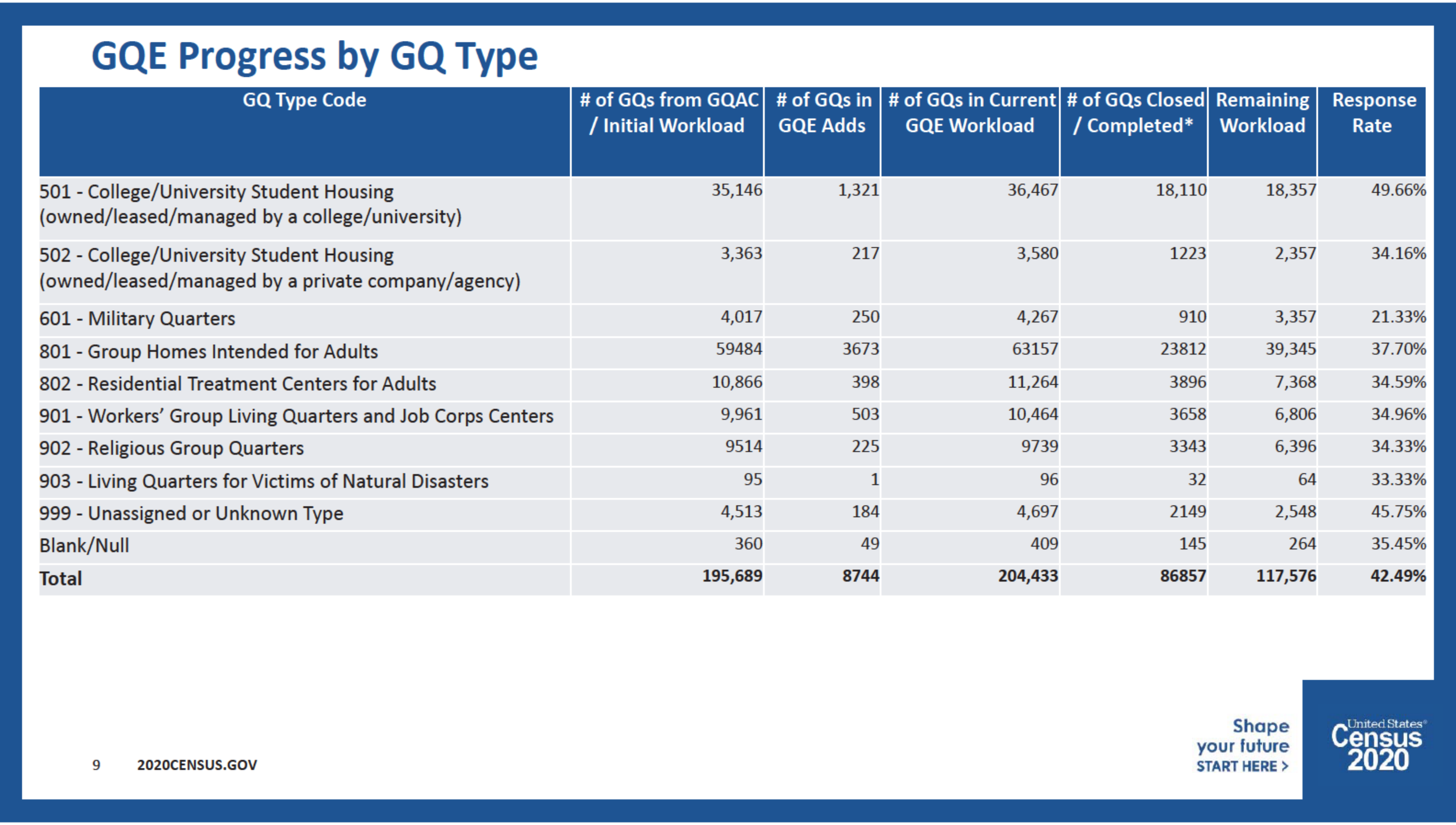
GQE Progress by GQ Type

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Current GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
101 - Federal Detention Centers	2,352	17	2,369	2,341	28	**98.82%**
102 - Federal Prisons	220	16	236	233	3	**98.73%**
103 - State Prisons	8,906	871	9,777	7,872	1,905	**80.52%**
104 - Local Jails and Other Municipal Confinement Facilities	3,707	91	3,798	1458	2,340	**38.39%**
105 - Correctional Residential Facilities	1,143	40	1,183	636	547	**53.76%**
106 - Military Disciplinary Barracks and Jails	38	0	38	13	25	**34.21%**
201 - Group Homes for Juveniles (non-correctional)	4,482	246	4,728	1,832	2,896	**38.75%**
202 - Residential Treatment Centers for Juveniles (non-correctional)	2,437	92	2,529	1124	1,405	**44.44%**
203 - Correctional Facilities Intended for Juveniles	1,902	23	1,925	967	958	**50.23%**
301 - Nursing Facilities/Skilled-Nursing Facilities	29,768	415	30,183	11,766	18,417	**38.98%**
401 - Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals	1,314	49	1,363	533	830	**39.10%**
402 - Hospitals with Patients Who Have No Usual Home Elsewhere	517	26	543	209	334	**38.49%**
403 - In-Patient Hospice Facilities	771	19	790	295	495	**37.34%**
404 - Military Treatment Facilities with Assigned Patients	37	0	37	14	23	**37.84%**
405 - Residential Schools for People with Disabilities	776	18	794	286	508	**36.02%**

Shape your future START HERE >

Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



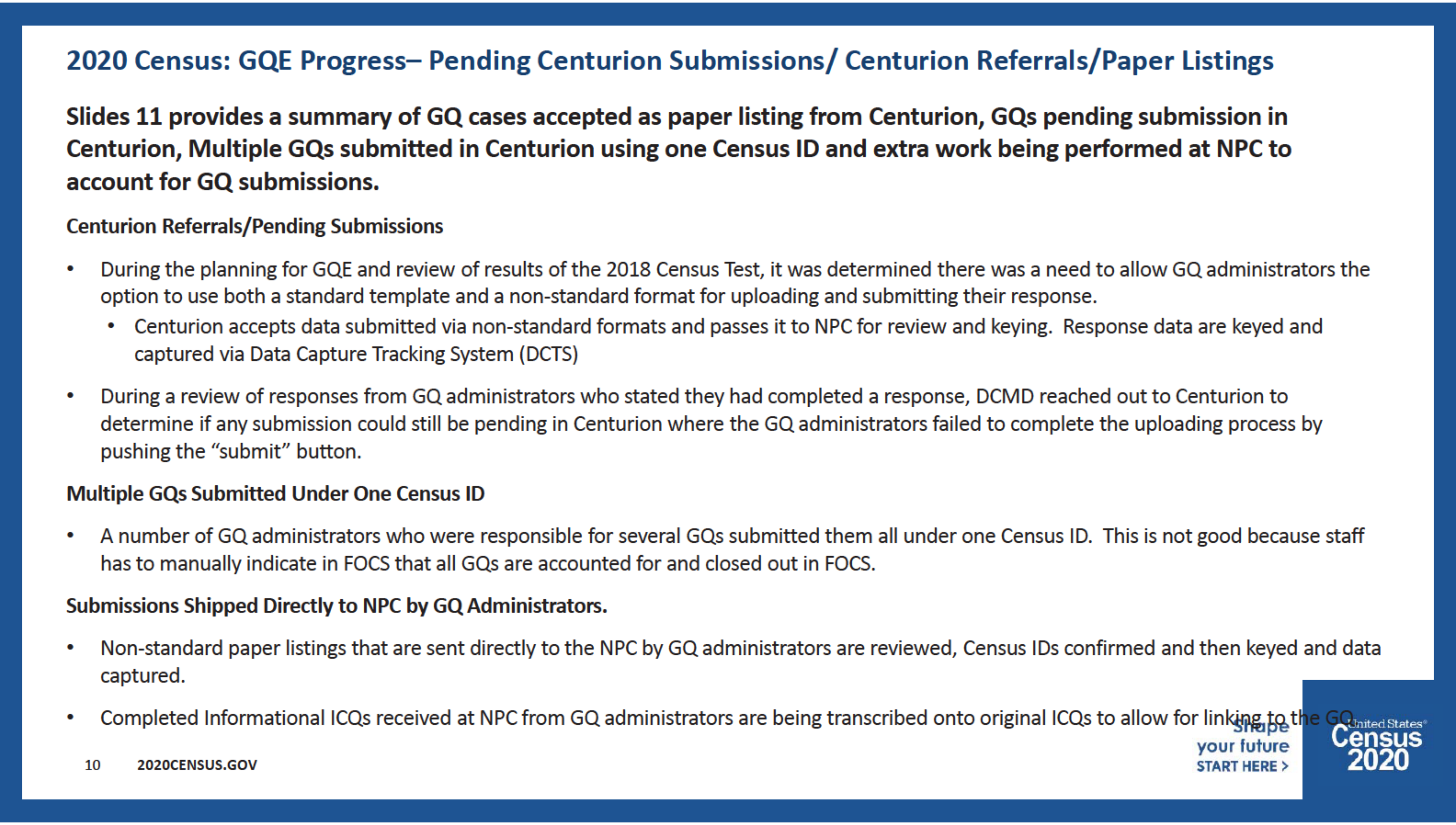
GQE Progress by GQ Type

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Current GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
501 - College/University Student Housing (owned/leased/managed by a college/university)	35,146	1,321	36,467	18,110	18,357	49.66%
502 - College/University Student Housing (owned/leased/managed by a private company/agency)	3,363	217	3,580	1223	2,357	34.16%
601 - Military Quarters	4,017	250	4,267	910	3,357	21.33%
801 - Group Homes Intended for Adults	59484	3673	63157	23812	39,345	37.70%
802 - Residential Treatment Centers for Adults	10,866	398	11,264	3896	7,368	34.59%
901 - Workers' Group Living Quarters and Job Corps Centers	9,961	503	10,464	3658	6,806	34.96%
902 - Religious Group Quarters	9514	225	9739	3343	6,396	34.33%
903 - Living Quarters for Victims of Natural Disasters	95	1	96	32	64	33.33%
999 - Unassigned or Unknown Type	4,513	184	4,697	2149	2,548	45.75%
Blank/Null	360	49	409	145	264	35.45%
Total	**195,689**	**8744**	**204,433**	**86857**	**117,576**	**42.49%**

Shape
your future
START HERE >

United States®
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census: GQE Progress– Pending Centurion Submissions/ Centurion Referrals/Paper Listings

Slides 11 provides a summary of GQ cases accepted as paper listing from Centurion, GQs pending submission in Centurion, Multiple GQs submitted in Centurion using one Census ID and extra work being performed at NPC to account for GQ submissions.

Centurion Referrals/Pending Submissions

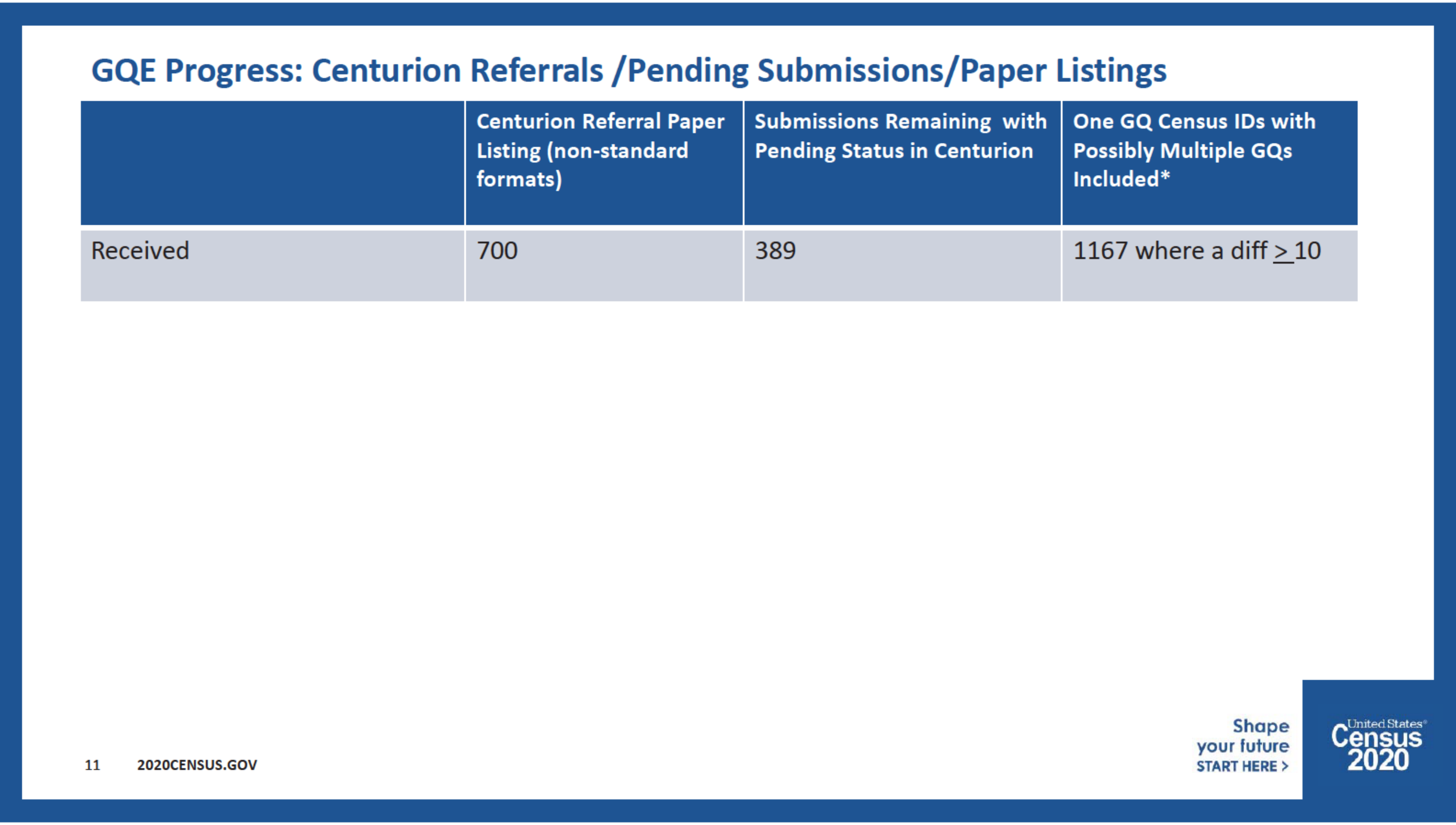
- During the planning for GQE and review of results of the 2018 Census Test, it was determined there was a need to allow GQ administrators the option to use both a standard template and a non-standard format for uploading and submitting their response.
  - Centurion accepts data submitted via non-standard formats and passes it to NPC for review and keying. Response data are keyed and captured via Data Capture Tracking System (DCTS)
- During a review of responses from GQ administrators who stated they had completed a response, DCMD reached out to Centurion to determine if any submission could still be pending in Centurion where the GQ administrators failed to complete the uploading process by pushing the "submit" button.

Multiple GQs Submitted Under One Census ID

- A number of GQ administrators who were responsible for several GQs submitted them all under one Census ID. This is not good because staff has to manually indicate in FOCS that all GQs are accounted for and closed out in FOCS.

Submissions Shipped Directly to NPC by GQ Administrators.

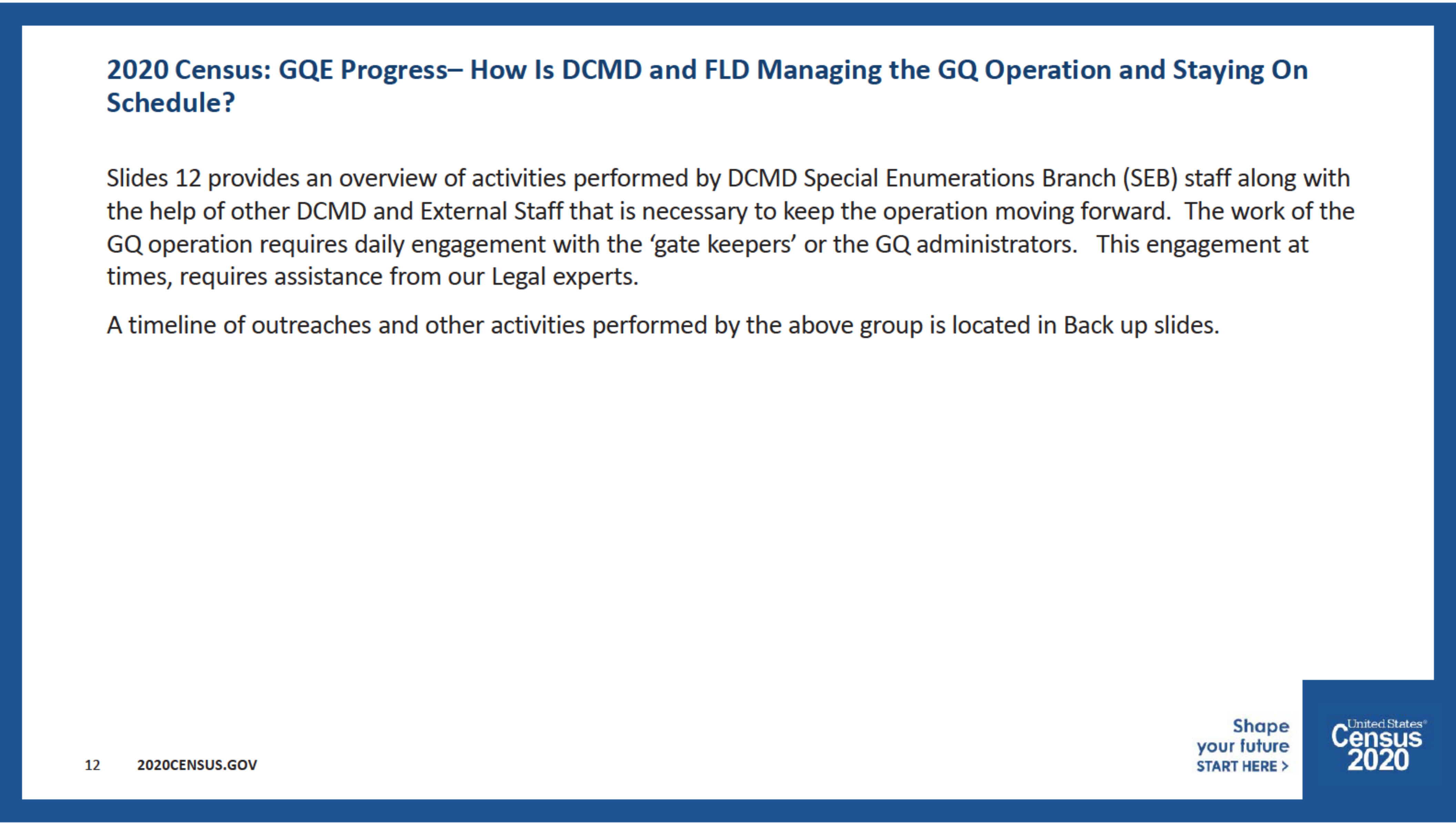
- Non-standard paper listings that are sent directly to the NPC by GQ administrators are reviewed, Census IDs confirmed and then keyed and data captured.
- Completed Informational ICQs received at NPC from GQ administrators are being transcribed onto original ICQs to allow for linking to the GQ



GQE Progress: Centurion Referrals /Pending Submissions/Paper Listings

	Centurion Referral Paper Listing (non-standard formats)	Submissions Remaining with Pending Status in Centurion	One GQ Census IDs with Possibly Multiple GQs Included*
Received	700	389	1167 where a diff ≥ 10

Shape your future START HERE >

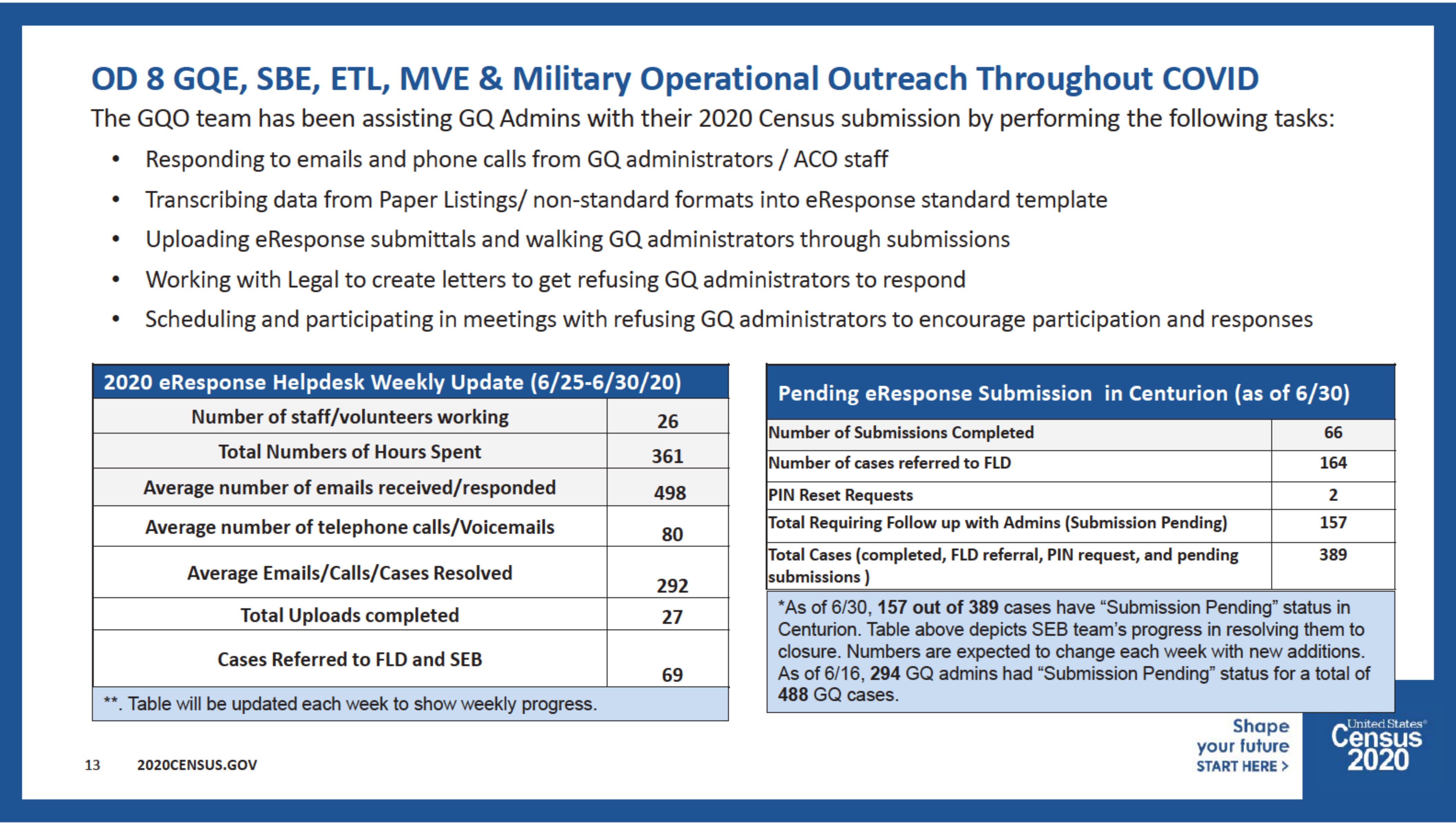


2020 Census: GQE Progress– How Is DCMD and FLD Managing the GQ Operation and Staying On Schedule?

Slides 12 provides an overview of activities performed by DCMD Special Enumerations Branch (SEB) staff along with the help of other DCMD and External Staff that is necessary to keep the operation moving forward. The work of the GQ operation requires daily engagement with the ‘gate keepers’ or the GQ administrators. This engagement at times, requires assistance from our Legal experts.

A timeline of outreaches and other activities performed by the above group is located in Back up slides.

DRB Approval Number: CBDRB-FY21-DSEP-002



OD 8 GQE, SBE, ETL, MVE & Military Operational Outreach Throughout COVID

The GQO team has been assisting GQ Admins with their 2020 Census submission by performing the following tasks:

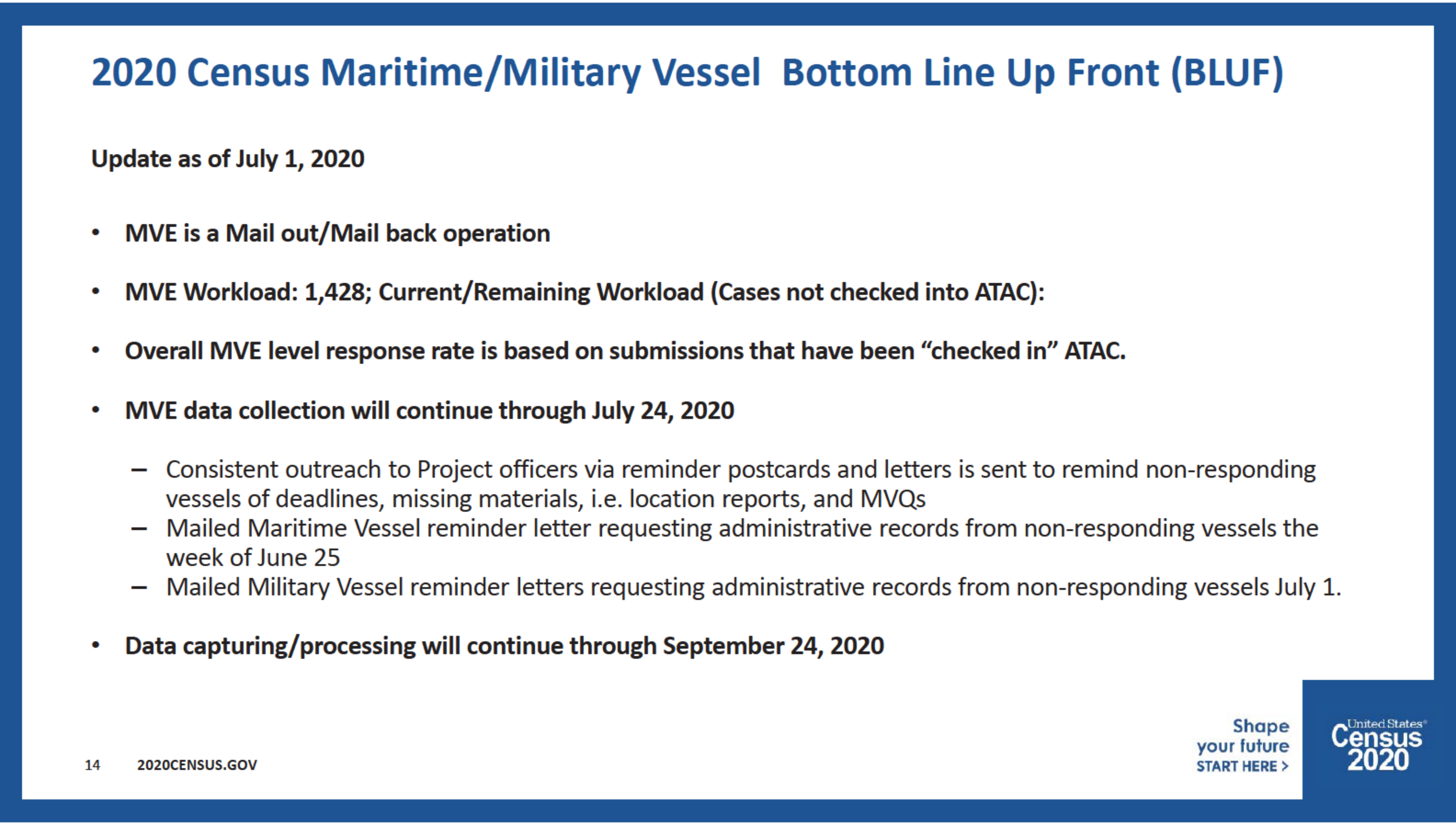
- Responding to emails and phone calls from GQ administrators / ACO staff
- Transcribing data from Paper Listings/ non-standard formats into eResponse standard template
- Uploading eResponse submittals and walking GQ administrators through submissions
- Working with Legal to create letters to get refusing GQ administrators to respond
- Scheduling and participating in meetings with refusing GQ administrators to encourage participation and responses

2020 eResponse Helpdesk Weekly Update (6/25-6/30/20)	
Number of staff/volunteers working	26
Total Numbers of Hours Spent	361
Average number of emails received/responded	498
Average number of telephone calls/Voicemails	80
Average Emails/Calls/Cases Resolved	292
Total Uploads completed	27
Cases Referred to FLD and SEB	69

**. Table will be updated each week to show weekly progress.

Pending eResponse Submission in Centurion (as of 6/30)	
Number of Submissions Completed	66
Number of cases referred to FLD	164
PIN Reset Requests	2
Total Requiring Follow up with Admins (Submission Pending)	157
Total Cases (completed, FLD referral, PIN request, and pending submissions)	389

*As of 6/30, **157 out of 389** cases have "Submission Pending" status in Centurion. Table above depicts SEB team's progress in resolving them to closure. Numbers are expected to change each week with new additions. As of 6/16, **294** GQ admins had "Submission Pending" status for a total of **488** GQ cases.



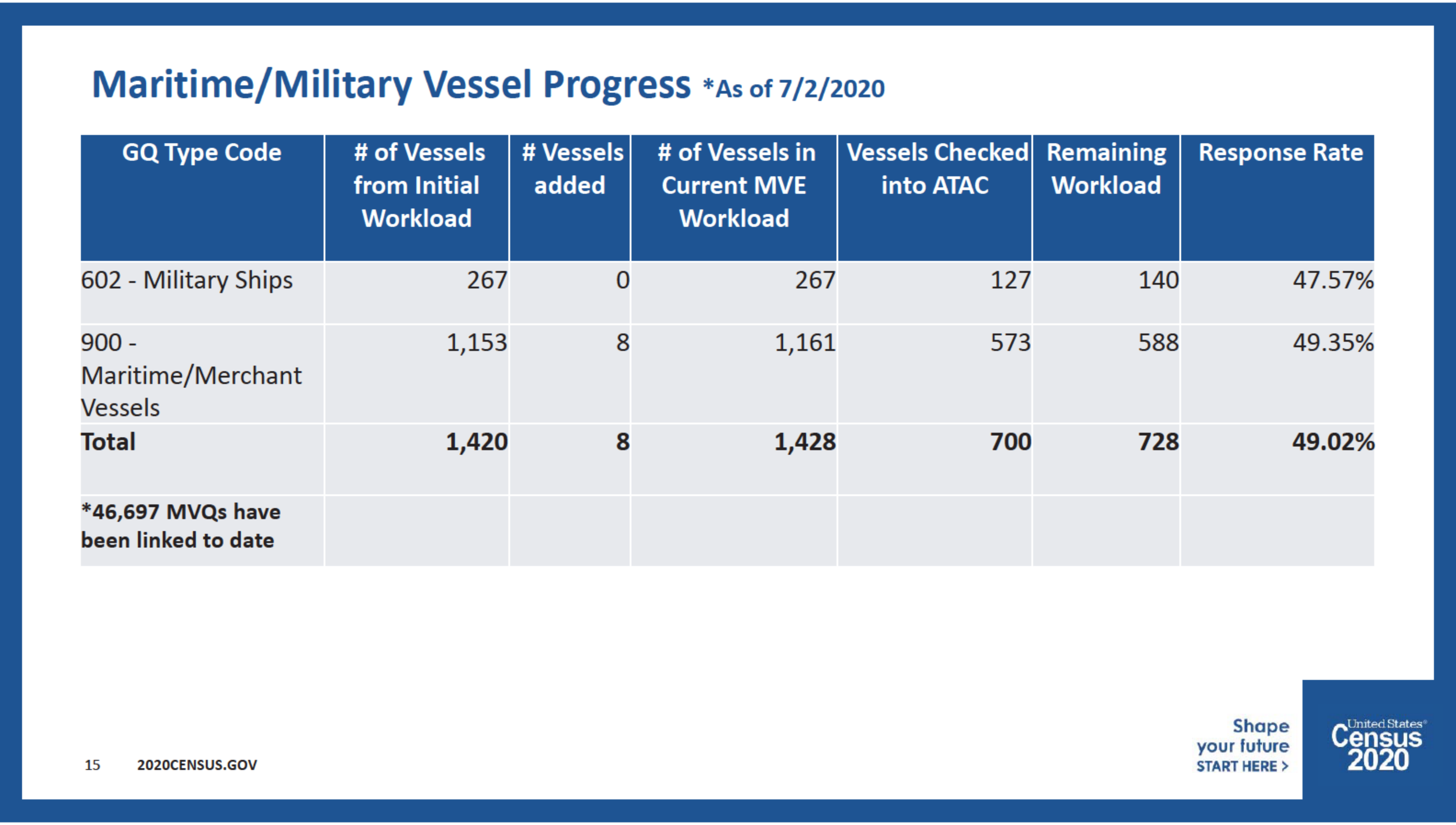
2020 Census Maritime/Military Vessel Bottom Line Up Front (BLUF)

Update as of July 1, 2020

- **MVE is a Mail out/Mail back operation**
- **MVE Workload: 1,428; Current/Remaining Workload (Cases not checked into ATAC):**
- **Overall MVE level response rate is based on submissions that have been "checked in" ATAC.**
- **MVE data collection will continue through July 24, 2020**
  - Consistent outreach to Project officers via reminder postcards and letters is sent to remind non-responding vessels of deadlines, missing materials, i.e. location reports, and MVQs
  - Mailed Maritime Vessel reminder letter requesting administrative records from non-responding vessels the week of June 25
  - Mailed Military Vessel reminder letters requesting administrative records from non-responding vessels July 1.
- **Data capturing/processing will continue through September 24, 2020**

Shape your future START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002

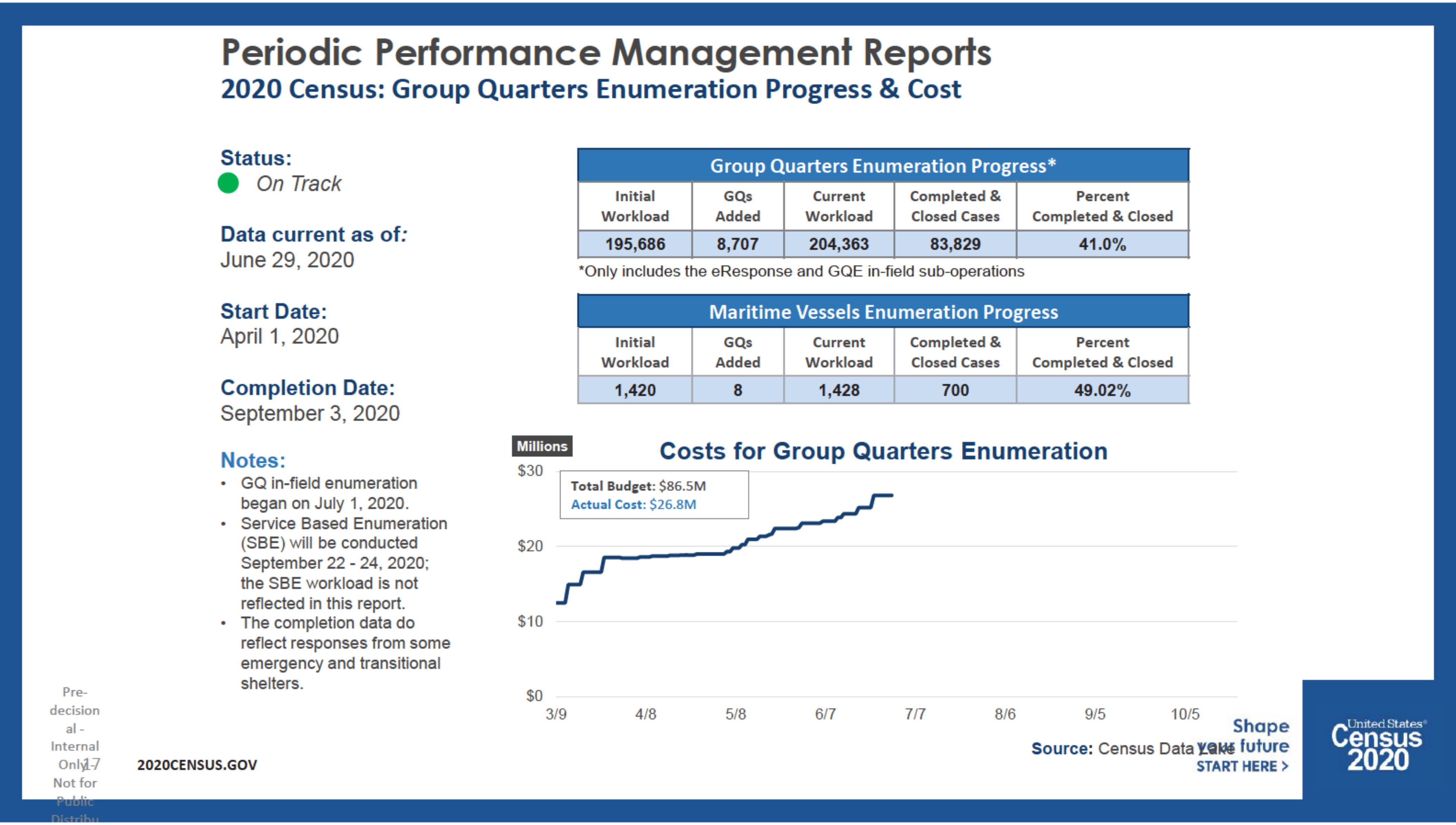


Maritime/Military Vessel Progress *As of 7/2/2020

GQ Type Code	# of Vessels from Initial Workload	# Vessels added	# of Vessels in Current MVE Workload	Vessels Checked into ATAC	Remaining Workload	Response Rate
602 - Military Ships	267	0	267	127	140	47.57%
900 - Maritime/Merchant Vessels	1,153	8	1,161	573	588	49.35%
Total	**1,420**	**8**	**1,428**	**700**	**728**	**49.02%**
***46,697 MVQs have been linked to date**						

Backup

Schedules and Systems



Periodic Performance Management Reports

2020 Census: Group Quarters Enumeration Progress & Cost

Status:
● *On Track*

Data current as of:
June 29, 2020

Start Date:
April 1, 2020

Completion Date:
September 3, 2020

Notes:

- GQ in-field enumeration began on July 1, 2020.
- Service Based Enumeration (SBE) will be conducted September 22 - 24, 2020; the SBE workload is not reflected in this report.
- The completion data do reflect responses from some emergency and transitional shelters.

Group Quarters Enumeration Progress*

Initial Workload	GQs Added	Current Workload	Completed & Closed Cases	Percent Completed & Closed
195,686	8,707	204,363	83,829	41.0%

*Only includes the eResponse and GQE in-field sub-operations

Maritime Vessels Enumeration Progress

Initial Workload	GQs Added	Current Workload	Completed & Closed Cases	Percent Completed & Closed
1,420	8	1,428	700	49.02%

Costs for Group Quarters Enumeration

Total Budget: $86.5M
Actual Cost: $26.8M

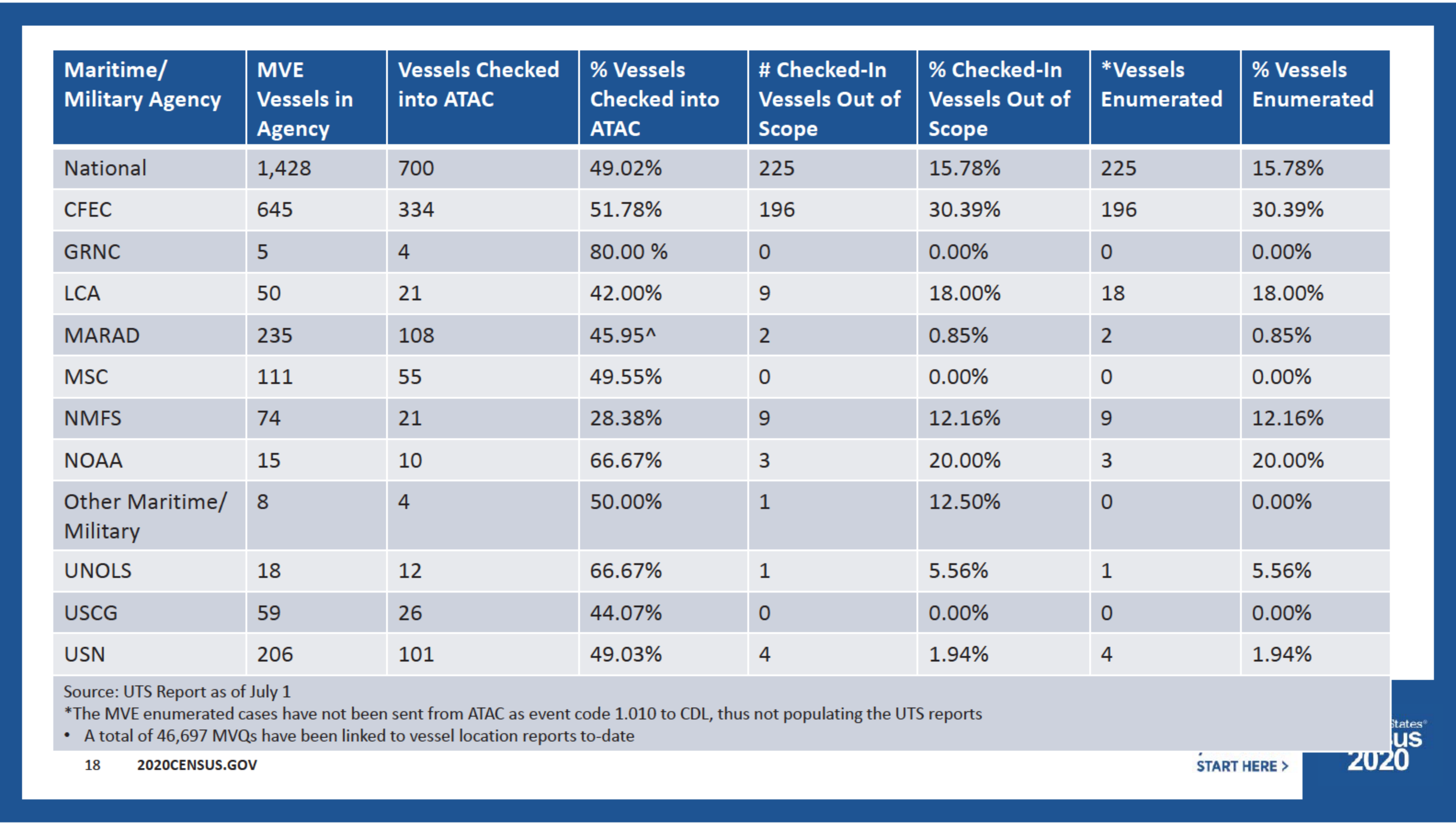
Source: Census Data Lake

2020CENSUS.GOV

Shape your future
START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



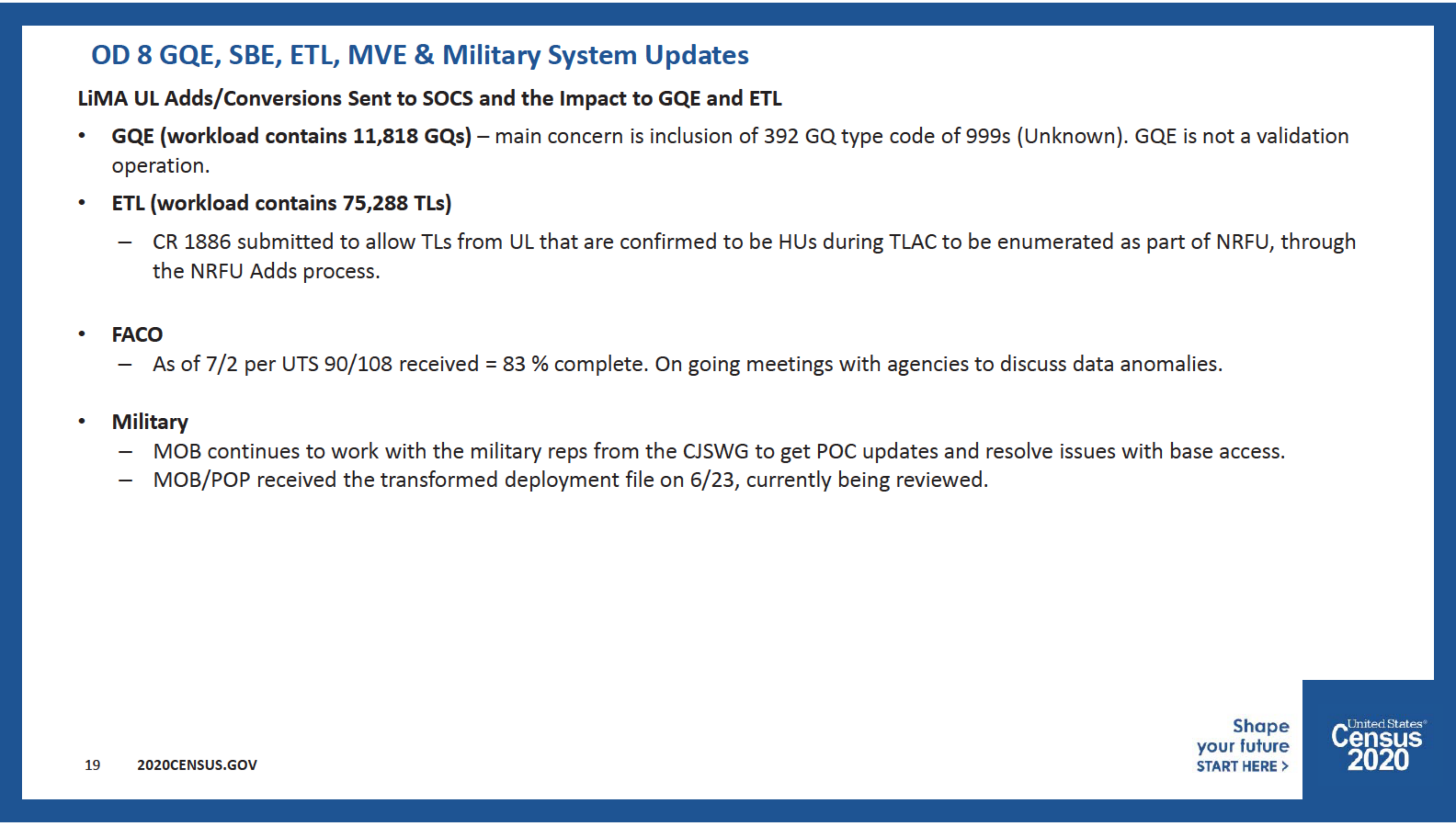
Maritime/ Military Agency	MVE Vessels in Agency	Vessels Checked into ATAC	% Vessels Checked into ATAC	# Checked-In Vessels Out of Scope	% Checked-In Vessels Out of Scope	*Vessels Enumerated	% Vessels Enumerated
National	1,428	700	49.02%	225	15.78%	225	15.78%
CFEC	645	334	51.78%	196	30.39%	196	30.39%
GRNC	5	4	80.00 %	0	0.00%	0	0.00%
LCA	50	21	42.00%	9	18.00%	18	18.00%
MARAD	235	108	45.95^	2	0.85%	2	0.85%
MSC	111	55	49.55%	0	0.00%	0	0.00%
NMFS	74	21	28.38%	9	12.16%	9	12.16%
NOAA	15	10	66.67%	3	20.00%	3	20.00%
Other Maritime/ Military	8	4	50.00%	1	12.50%	0	0.00%
UNOLS	18	12	66.67%	1	5.56%	1	5.56%
USCG	59	26	44.07%	0	0.00%	0	0.00%
USN	206	101	49.03%	4	1.94%	4	1.94%

Source: UTS Report as of July 1

*The MVE enumerated cases have not been sent from ATAC as event code 1.010 to CDL, thus not populating the UTS reports

- A total of 46,697 MVQs have been linked to vessel location reports to-date

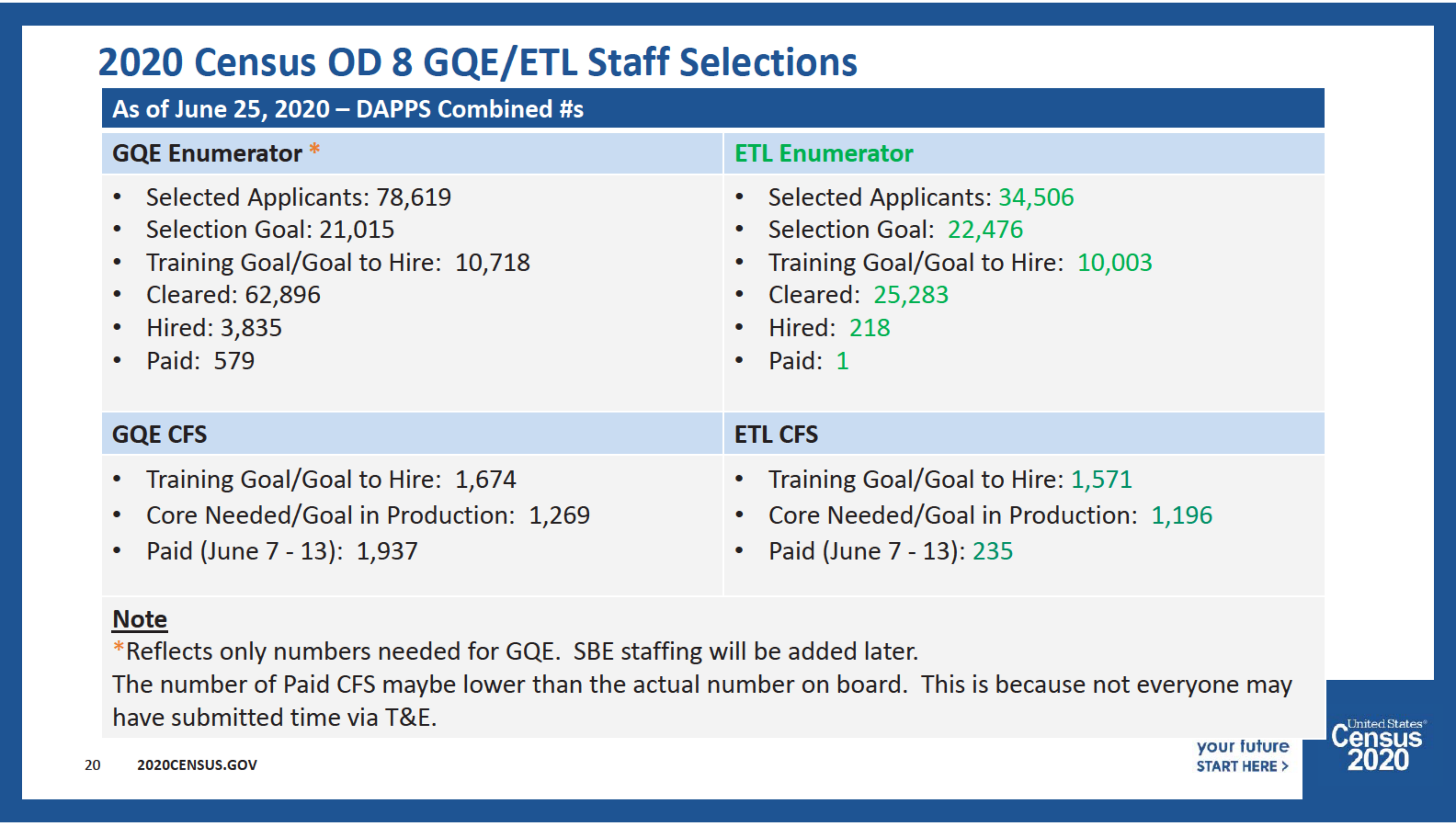
START HERE >



OD 8 GQE, SBE, ETL, MVE & Military System Updates

LiMA UL Adds/Conversions Sent to SOCS and the Impact to GQE and ETL

- **GQE (workload contains 11,818 GQs)** – main concern is inclusion of 392 GQ type code of 999s (Unknown). GQE is not a validation operation.
- **ETL (workload contains 75,288 TLs)**
  - CR 1886 submitted to allow TLs from UL that are confirmed to be HUs during TLAC to be enumerated as part of NRFU, through the NRFU Adds process.
- **FACO**
  - As of 7/2 per UTS 90/108 received = 83 % complete. On going meetings with agencies to discuss data anomalies.
- **Military**
  - MOB continues to work with the military reps from the CJSWG to get POC updates and resolve issues with base access.
  - MOB/POP received the transformed deployment file on 6/23, currently being reviewed.



2020 Census OD 8 GQE/ETL Staff Selections

As of June 25, 2020 – DAPPS Combined #s

GQE Enumerator *

- Selected Applicants: 78,619
- Selection Goal: 21,015
- Training Goal/Goal to Hire: 10,718
- Cleared: 62,896
- Hired: 3,835
- Paid: 579

ETL Enumerator

- Selected Applicants: 34,506
- Selection Goal: 22,476
- Training Goal/Goal to Hire: 10,003
- Cleared: 25,283
- Hired: 218
- Paid: 1

GQE CFS

- Training Goal/Goal to Hire: 1,674
- Core Needed/Goal in Production: 1,269
- Paid (June 7 - 13): 1,937

ETL CFS

- Training Goal/Goal to Hire: 1,571
- Core Needed/Goal in Production: 1,196
- Paid (June 7 - 13): 235

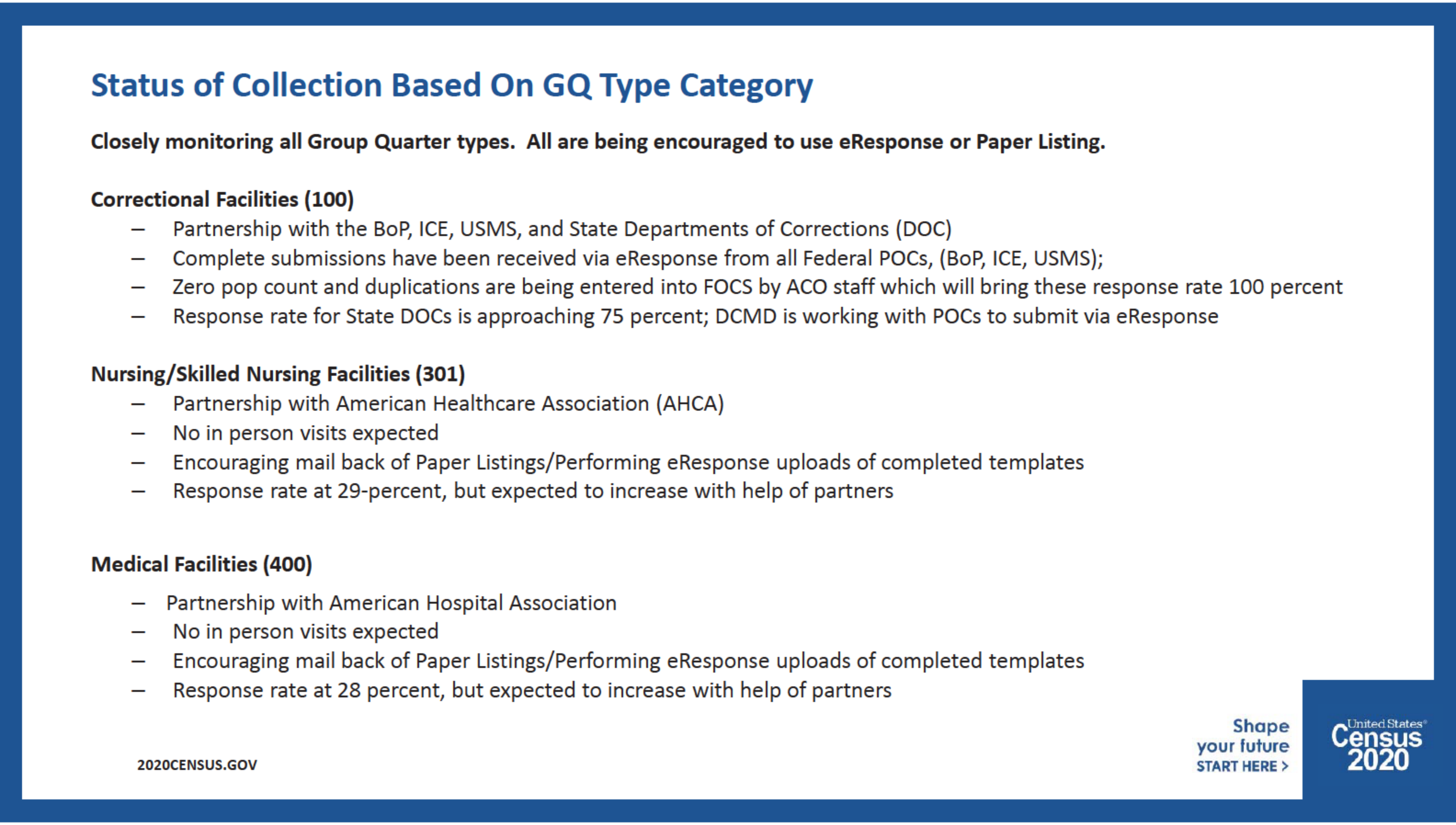
Note

*Reflects only numbers needed for GQE. SBE staffing will be added later.

The number of Paid CFS maybe lower than the actual number on board. This is because not everyone may have submitted time via T&E.

your future START HERE >

United States® Census 2020



Status of Collection Based On GQ Type Category

Closely monitoring all Group Quarter types. All are being encouraged to use eResponse or Paper Listing.

Correctional Facilities (100)

– Partnership with the BoP, ICE, USMS, and State Departments of Corrections (DOC)
– Complete submissions have been received via eResponse from all Federal POCs, (BoP, ICE, USMS);
– Zero pop count and duplications are being entered into FOCS by ACO staff which will bring these response rate 100 percent
– Response rate for State DOCs is approaching 75 percent; DCMD is working with POCs to submit via eResponse

Nursing/Skilled Nursing Facilities (301)

– Partnership with American Healthcare Association (AHCA)
– No in person visits expected
– Encouraging mail back of Paper Listings/Performing eResponse uploads of completed templates
– Response rate at 29-percent, but expected to increase with help of partners

Medical Facilities (400)

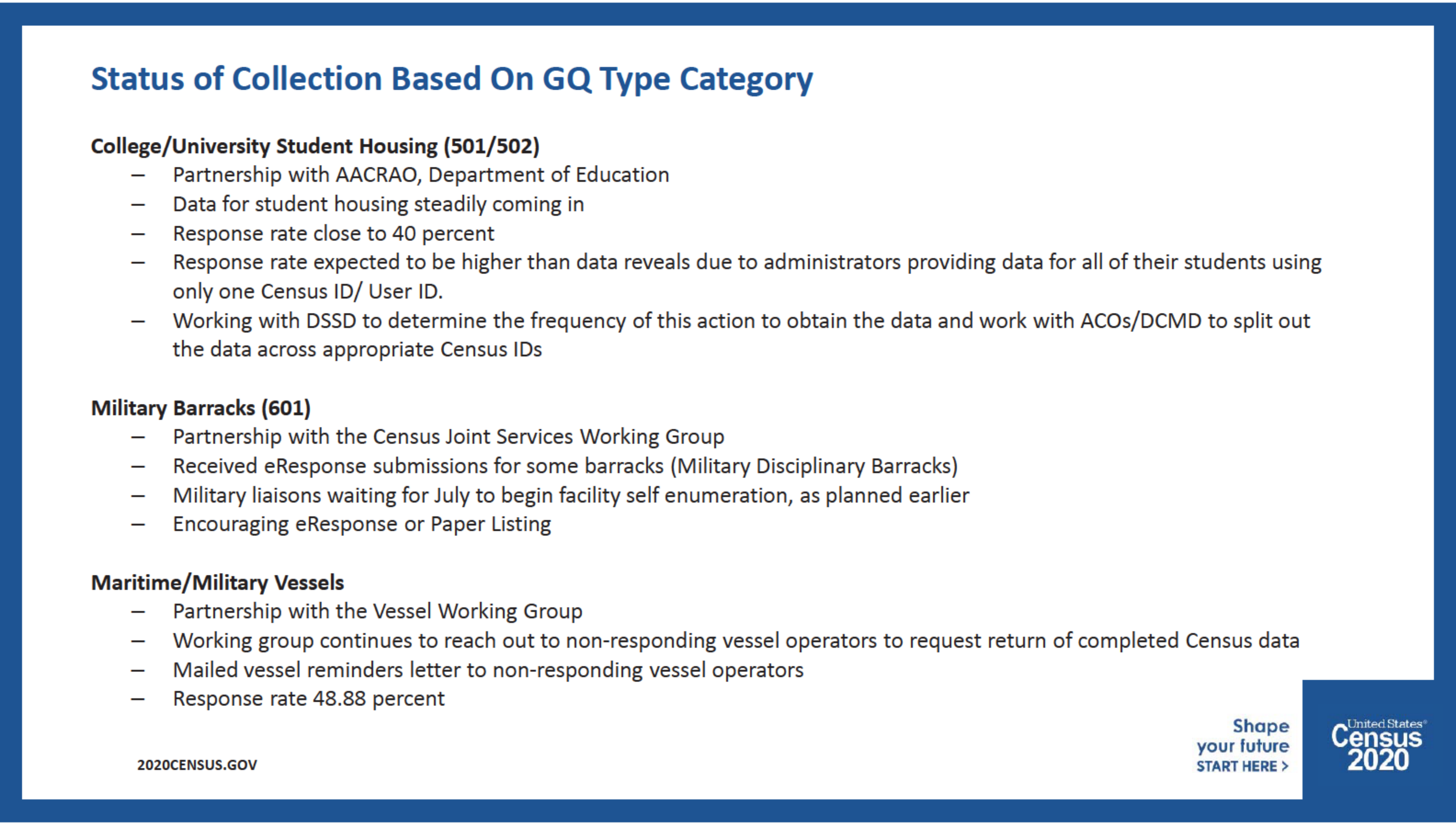
– Partnership with American Hospital Association
– No in person visits expected
– Encouraging mail back of Paper Listings/Performing eResponse uploads of completed templates
– Response rate at 28 percent, but expected to increase with help of partners

2020CENSUS.GOV

Shape
your future
START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



Status of Collection Based On GQ Type Category

College/University Student Housing (501/502)

- Partnership with AACRAO, Department of Education
- Data for student housing steadily coming in
- Response rate close to 40 percent
- Response rate expected to be higher than data reveals due to administrators providing data for all of their students using only one Census ID/ User ID.
- Working with DSSD to determine the frequency of this action to obtain the data and work with ACOs/DCMD to split out the data across appropriate Census IDs

Military Barracks (601)

- Partnership with the Census Joint Services Working Group
- Received eResponse submissions for some barracks (Military Disciplinary Barracks)
- Military liaisons waiting for July to begin facility self enumeration, as planned earlier
- Encouraging eResponse or Paper Listing

Maritime/Military Vessels

- Partnership with the Vessel Working Group
- Working group continues to reach out to non-responding vessel operators to request return of completed Census data
- Mailed vessel reminders letter to non-responding vessel operators
- Response rate 48.88 percent

2020CENSUS.GOV

Shape
your future
START HERE >

United States®
Census
2020

Status of Collection Based On GQ Type Category

Group Homes/Residential Treatment Centers (Juvenile/Adults) (200/800)

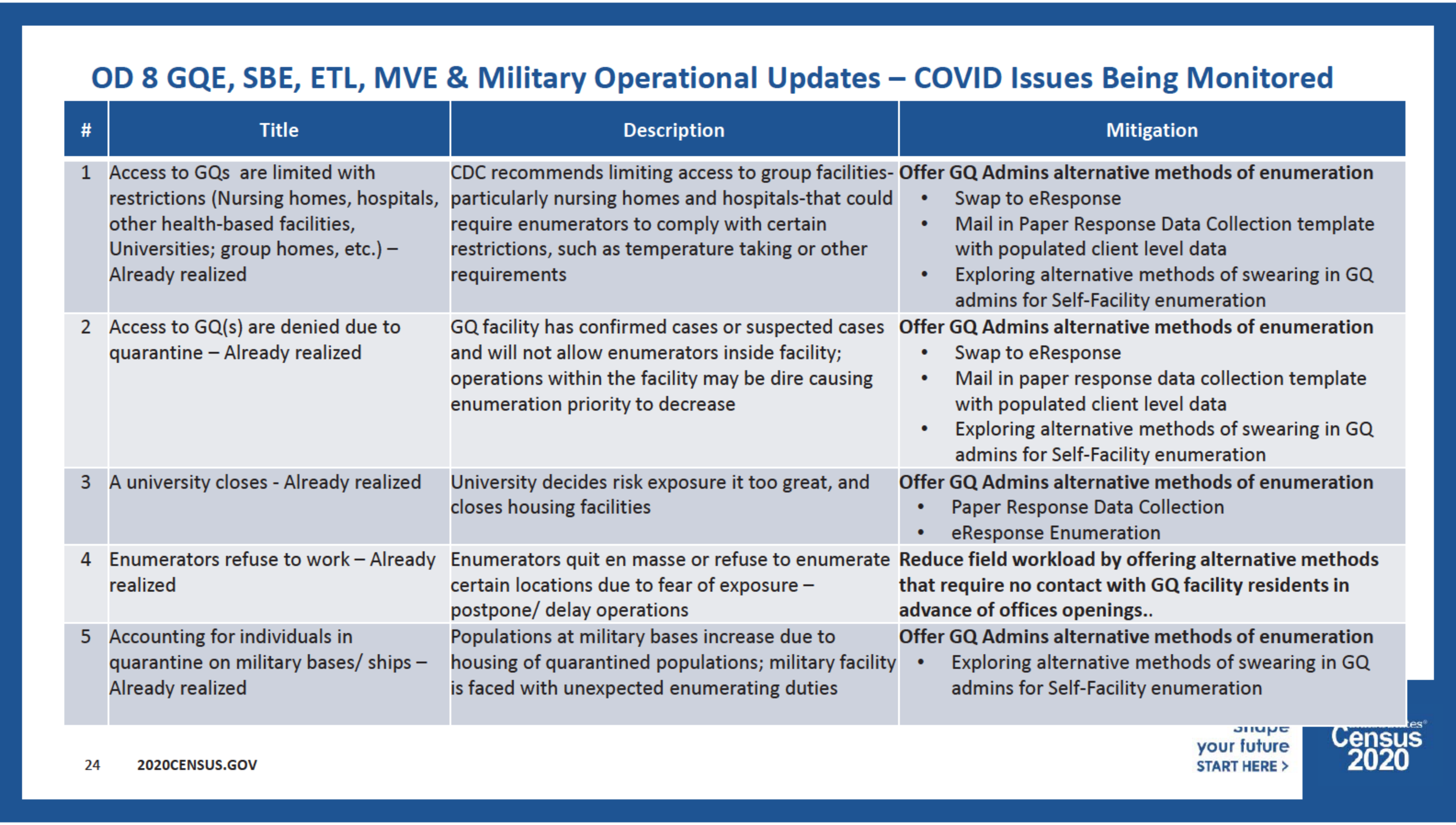
- Worked with Legal to obtain a standard HIPAA compliance letter to share with juvenile and treatment facilities concerned about releasing health care information
- Includes language received from the Department of Health and Human Services regarding the Privacy Rule, under the Health Insurance Portability and Accountability Act of 1996 (HIPPA), which addresses the use and disclosure of individual health information. The Privacy Rule also sets standards for individuals' privacy rights, to understand and control how their health information is used. Since the decennial census does not request any health information, neither HIPPA nor the Privacy Rule bars response.
- Assisting GQ administrators with data uploads to minimize level of effort and stress during this pandemic

Service-Based Facilities (700)

- Three day In-person enumeration to began 9/22

All GQ types with the exception of SBEs are being encouraged to use eResponse or Paper Listing and closely monitored.

DRB Approval Number: CBDRB-FY21-DSEP-002

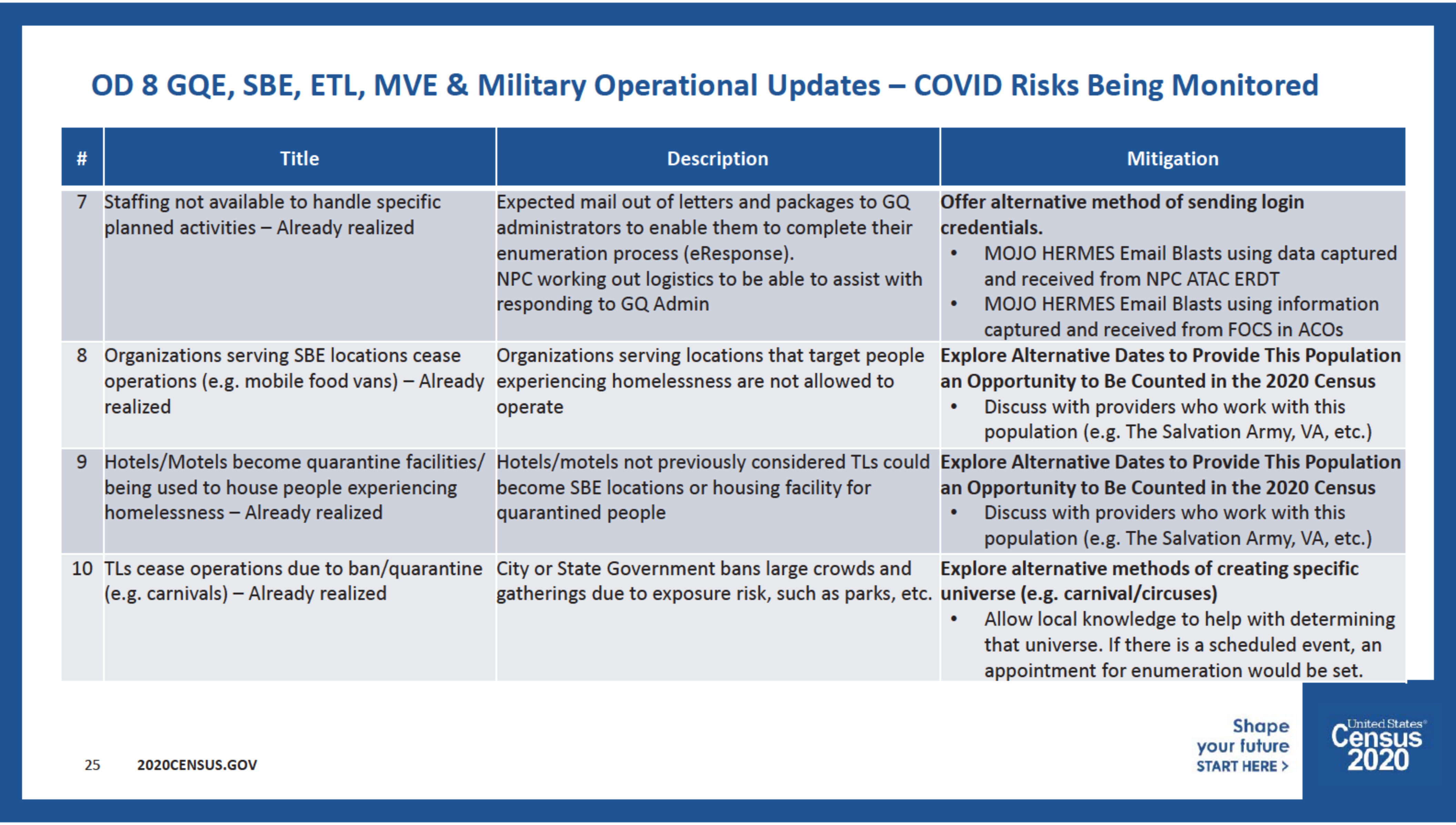


OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Issues Being Monitored

#	Title	Description	Mitigation
1	Access to GQs are limited with restrictions (Nursing homes, hospitals, other health-based facilities, Universities; group homes, etc.) – Already realized	CDC recommends limiting access to group facilities- particularly nursing homes and hospitals-that could require enumerators to comply with certain restrictions, such as temperature taking or other requirements	**Offer GQ Admins alternative methods of enumeration** • Swap to eResponse • Mail in Paper Response Data Collection template with populated client level data • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration
2	Access to GQ(s) are denied due to quarantine – Already realized	GQ facility has confirmed cases or suspected cases and will not allow enumerators inside facility; operations within the facility may be dire causing enumeration priority to decrease	**Offer GQ Admins alternative methods of enumeration** • Swap to eResponse • Mail in paper response data collection template with populated client level data • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration
3	A university closes - Already realized	University decides risk exposure it too great, and closes housing facilities	**Offer GQ Admins alternative methods of enumeration** • Paper Response Data Collection • eResponse Enumeration
4	Enumerators refuse to work – Already realized	Enumerators quit en masse or refuse to enumerate certain locations due to fear of exposure – postpone/ delay operations	**Reduce field workload by offering alternative methods that require no contact with GQ facility residents in advance of offices openings..**
5	Accounting for individuals in quarantine on military bases/ ships – Already realized	Populations at military bases increase due to housing of quarantined populations; military facility is faced with unexpected enumerating duties	**Offer GQ Admins alternative methods of enumeration** • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration

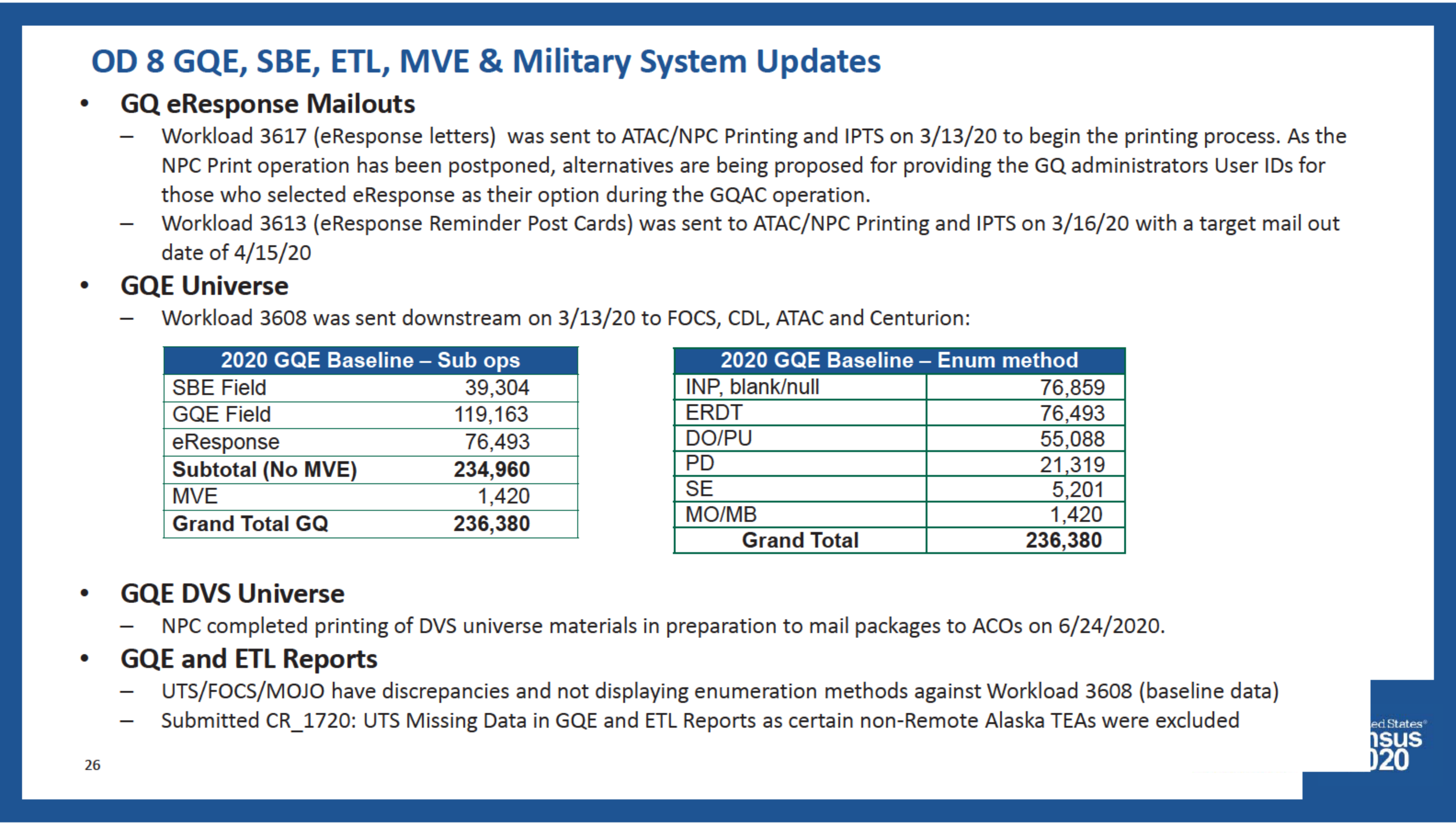
Shape your future START HERE >

United States Census 2020



OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Risks Being Monitored

#	Title	Description	Mitigation
7	Staffing not available to handle specific planned activities – Already realized	Expected mail out of letters and packages to GQ administrators to enable them to complete their enumeration process (eResponse). NPC working out logistics to be able to assist with responding to GQ Admin	**Offer alternative method of sending login credentials.** • MOJO HERMES Email Blasts using data captured and received from NPC ATAC ERDT • MOJO HERMES Email Blasts using information captured and received from FOCS in ACOs
8	Organizations serving SBE locations cease operations (e.g. mobile food vans) – Already realized	Organizations serving locations that target people experiencing homelessness are not allowed to operate	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
9	Hotels/Motels become quarantine facilities/ being used to house people experiencing homelessness – Already realized	Hotels/motels not previously considered TLs could become SBE locations or housing facility for quarantined people	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
10	TLs cease operations due to ban/quarantine (e.g. carnivals) – Already realized	City or State Government bans large crowds and gatherings due to exposure risk, such as parks, etc.	**Explore alternative methods of creating specific universe (e.g. carnival/circuses)** • Allow local knowledge to help with determining that universe. If there is a scheduled event, an appointment for enumeration would be set.



OD 8 GQE, SBE, ETL, MVE & Military System Updates

- **GQ eResponse Mailouts**
  - Workload 3617 (eResponse letters) was sent to ATAC/NPC Printing and IPTS on 3/13/20 to begin the printing process. As the NPC Print operation has been postponed, alternatives are being proposed for providing the GQ administrators User IDs for those who selected eResponse as their option during the GQAC operation.
  - Workload 3613 (eResponse Reminder Post Cards) was sent to ATAC/NPC Printing and IPTS on 3/16/20 with a target mail out date of 4/15/20
- **GQE Universe**
  - Workload 3608 was sent downstream on 3/13/20 to FOCS, CDL, ATAC and Centurion:

2020 GQE Baseline – Sub ops	
SBE Field	39,304
GQE Field	119,163
eResponse	76,493
Subtotal (No MVE)	**234,960**
MVE	1,420
Grand Total GQ	**236,380**

2020 GQE Baseline – Enum method	
INP, blank/null	76,859
ERDT	76,493
DO/PU	55,088
PD	21,319
SE	5,201
MO/MB	1,420
Grand Total	**236,380**

- **GQE DVS Universe**
  - NPC completed printing of DVS universe materials in preparation to mail packages to ACOs on 6/24/2020.
- **GQE and ETL Reports**
  - UTS/FOCS/MOJO have discrepancies and not displaying enumeration methods against Workload 3608 (baseline data)
  - Submitted CR_1720: UTS Missing Data in GQE and ETL Reports as certain non-Remote Alaska TEAs were excluded

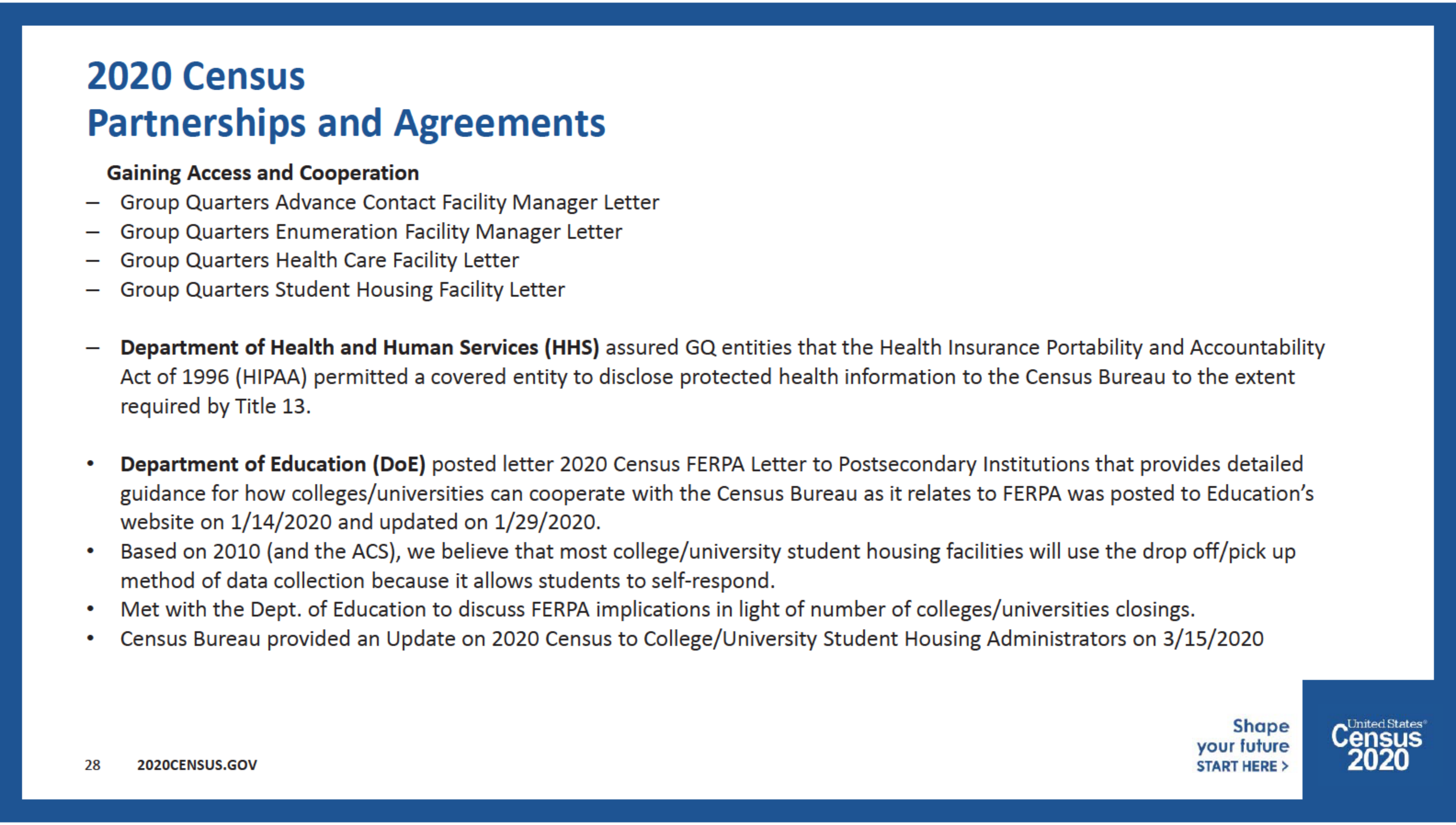
OD 8 GQE, SBE, ETL, MVE & Military Operational Outreach Throughout COVID

- March – Nationwide "Stay at Home" orders
- **March:** Letter developed by DCMD for RCCs to send to GQs that selected self enumeration options to change option to eResponse or paper listings
- **March 13:** Posted letter, *Update on 2020 Census for Student Housing Administrators* on the Census Bureau and Department of Education Website, requesting administrators who selected self enumeration option to change option to eResponse or Paper Listings
- **March 25:** DCMD sent *Update on the 2020 Census for Health Care Administrator* letter to Health Care umbrella organizations providing guidance for Administrators that selected a self-enumeration option to change method of enumeration to eResponse of Paper Listings
- NPC/ Jeffersonville call center was closed due to COVID and was not available to complete task in support of eResponse. As a result:
  - NPC was unable to meet the March 27 deadline for mailing eResponse Letters with Login credentials
  - DCMD worked with NPC ATAC management to update the system to allow multiple users to view and update email address
  - DCMD stood up a Call Center with staff across ACOs and Census HQ volunteers to verify/update email address to deliver login credentials.
- **March 31:** Mojo/Hermes sent out 1st email blast with login credentials to GQ admins who has selected eResponse during GQAC
- **April 1:** GQE eResponse portal became available for GQ submittals.
- **April 13** and **April 20:** Mojo/Hermes sent out email blast 2 and 3 with login credentials for bounce back emails from 1st email blast
- **April 2:** DCMD Staff and volunteers across the Decennial Directorate and other started reaching out in response to questions from GQ administrators received via email and phone calls.
- April ? PIO developed a video to college students. Posted on website encouraging internet response or that GQ admins would respond for them if they live in student housing
- **April 20:** ACO began calling GQ admins to offer Mail out/ Mail back Paper Response Data Collection (Paper Listings)
- **May 28:** Census Bureau participated in a webinar to remind/ update student housing administrators on the 2020 Census Group Quarters operation and to inform administrators of the upcoming request for off-campus student data.
- **June 3:** NPC mailed Maritime/Military Vessel reminder letters to non-responding vessel operators
- **June 8:** ACO staff began calling GQ administrators to reschedule appointment dates for their facilities.
- **June 11:** Meeting with AACRAO…
- **June 22:** Provided updates to The Salvation Army for their Directive to be sent to their managing entities.
- **June 22:** Met with National Network to End Domestic Violence to discuss upcoming enumeration, options, and COVID-19 procedures

Along

Shape your future START HERE >

United States Census 2020



2020 Census
Partnerships and Agreements

Gaining Access and Cooperation

- Group Quarters Advance Contact Facility Manager Letter
- Group Quarters Enumeration Facility Manager Letter
- Group Quarters Health Care Facility Letter
- Group Quarters Student Housing Facility Letter

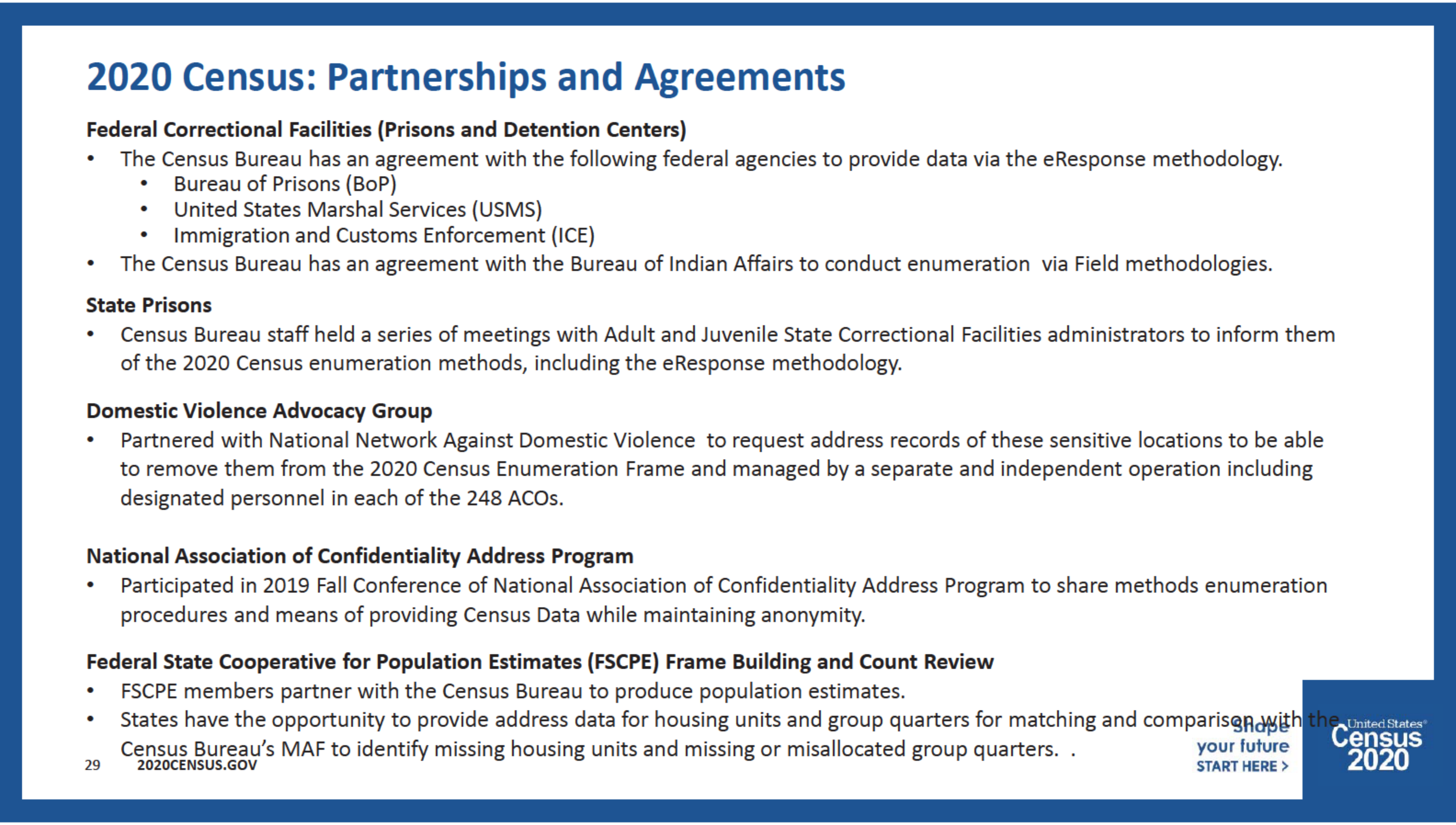
- **Department of Health and Human Services (HHS)** assured GQ entities that the Health Insurance Portability and Accountability Act of 1996 (HIPAA) permitted a covered entity to disclose protected health information to the Census Bureau to the extent required by Title 13.

- **Department of Education (DoE)** posted letter 2020 Census FERPA Letter to Postsecondary Institutions that provides detailed guidance for how colleges/universities can cooperate with the Census Bureau as it relates to FERPA was posted to Education's website on 1/14/2020 and updated on 1/29/2020.
- Based on 2010 (and the ACS), we believe that most college/university student housing facilities will use the drop off/pick up method of data collection because it allows students to self-respond.
- Met with the Dept. of Education to discuss FERPA implications in light of number of colleges/universities closings.
- Census Bureau provided an Update on 2020 Census to College/University Student Housing Administrators on 3/15/2020

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census: Partnerships and Agreements

Federal Correctional Facilities (Prisons and Detention Centers)

- The Census Bureau has an agreement with the following federal agencies to provide data via the eResponse methodology.
  - Bureau of Prisons (BoP)
  - United States Marshal Services (USMS)
  - Immigration and Customs Enforcement (ICE)
- The Census Bureau has an agreement with the Bureau of Indian Affairs to conduct enumeration via Field methodologies.

State Prisons

- Census Bureau staff held a series of meetings with Adult and Juvenile State Correctional Facilities administrators to inform them of the 2020 Census enumeration methods, including the eResponse methodology.

Domestic Violence Advocacy Group

- Partnered with National Network Against Domestic Violence to request address records of these sensitive locations to be able to remove them from the 2020 Census Enumeration Frame and managed by a separate and independent operation including designated personnel in each of the 248 ACOs.

National Association of Confidentiality Address Program

- Participated in 2019 Fall Conference of National Association of Confidentiality Address Program to share methods enumeration procedures and means of providing Census Data while maintaining anonymity.

Federal State Cooperative for Population Estimates (FSCPE) Frame Building and Count Review

- FSCPE members partner with the Census Bureau to produce population estimates.
- States have the opportunity to provide address data for housing units and group quarters for matching and comparison with the Census Bureau's MAF to identify missing housing units and missing or misallocated group quarters. .

2020 Census
Partnerships and Agreements

Salvation Army

- The Salvation completed and will provide each of their entities with the appropriate Directive.
  - 2020 Decennial Population Census – General (Handling of Soup kitchens and mobile food vans)
  - 2020 Decennial Population Census – Residential Institutions (Adult Rehabilitation Centers, Harbor Light Centers, Transient Lodges, residential facilities for children and other temporary housing facilities such as shelters for men, women and/or families)

American Healthcare Association (AHCA)

- The Census Bureau provided final updates for the 2020 Regulatory Advisory Letter to be sent to entities of the AHCA and the National Center for Assisted Living (NCAL) to inform them of 2020 Census enumeration plans.
- The Census Bureau provided additional feedback to the 2020 Census Procedures Outlined for Long Term Care (LTC) Facilities from AHCA who also had questions about classifying their Intermediate Care Facilities for Individuals with Intellectual Disabilities Facilities (ICFs/ID) as either GQs and housing units and concerns about the health of their residents and census workers in light of the COVID-19 Outbreak. POP provided language on what health care facilities could expect, depending on which operation they had already been assigned to.

Veteran Affairs

- The Veteran Affairs/ Administration provided Census with address records of location where they house veterans who may be experiencing homelessness.

2020 Census
Partnerships and Agreements

American **Hospital** Association (**AHA**)

- Received Final AHA Member Advisory that has been emailed and place on the AHA website for AHA member access.

Department of **Defense**

- DoD signed the Letter of Support and Tasking Memo 11/13/19.

OD 8 GQE, SBE, ETL, MVE & Military Operational Updates

2020census.gov website

Conducting the Count: https://2020census.gov/en/conducting-the-count.html

Counting People in Group Living Arrangements: https://2020census.gov/en/conducting-the-count/gq.html

Group Quarters Enumeration: https://2020census.gov/en/conducting-the-count/gq/gqe.html

Service-Based Enumeration: https://2020census.gov/en/conducting-the-count/gq/sbe.html

Group Quarters Advance Contact: https://2020census.gov/en/conducting-the-count/gq/gqac.html

eResponse: https://2020census.gov/en/conducting-the-count/gq/eresponse.html

Maritime and Military Vessel Enumeration: https://2020census.gov/en/conducting-the-count/gq/mve.html

Department of Education Student Privacy Policy Office: https://studentprivacy.ed.gov/faq/colleges-and-2020-census

DRB Approval Number: CBDRB-FY21-DSEP-002

Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

United States Census 2020

More information on the 2020 Census: http://www.census.gov/2020Census

American Community Survey

More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

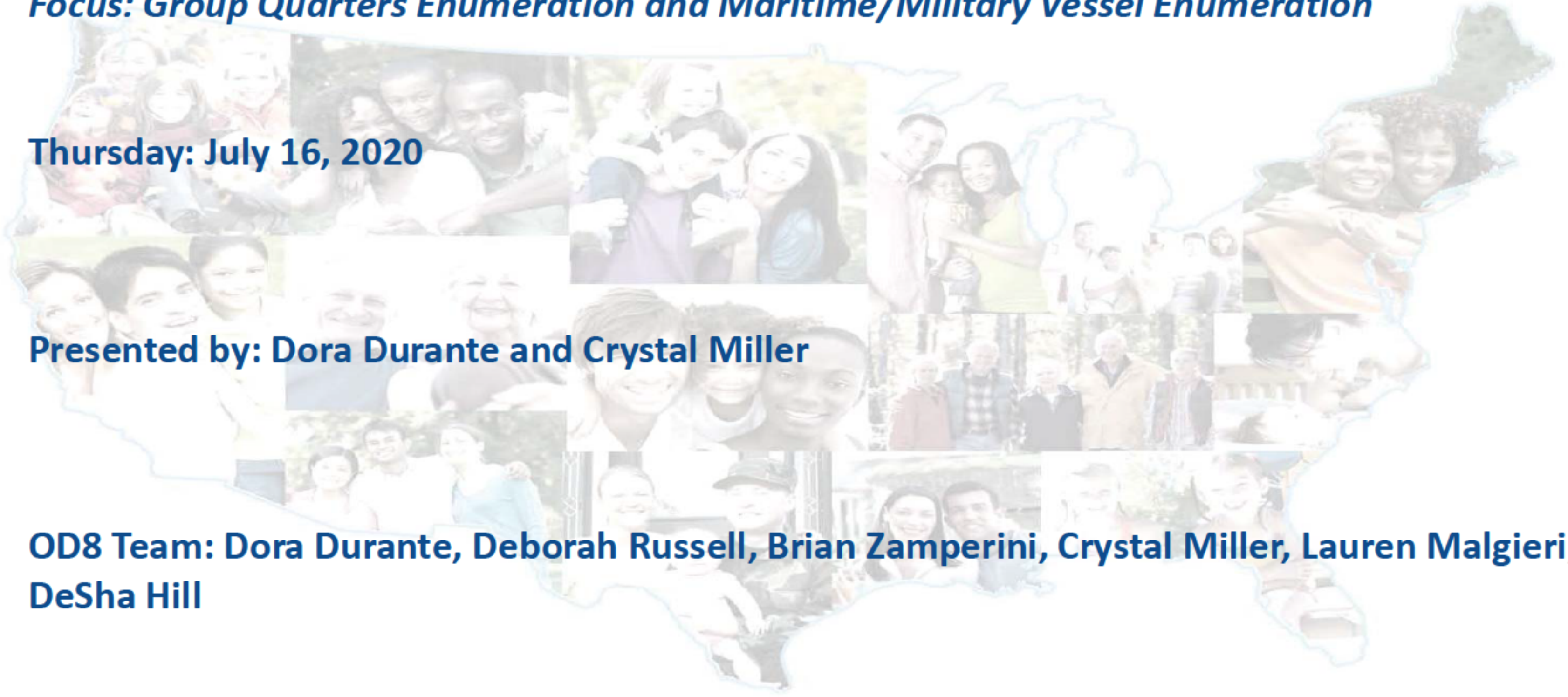
twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

instagram.com/uscensusbureau

2020 Census Operational Delivery 8: Group Quarters Update

Focus: Group Quarters Enumeration and Maritime/Military Vessel Enumeration



Thursday: July 16, 2020

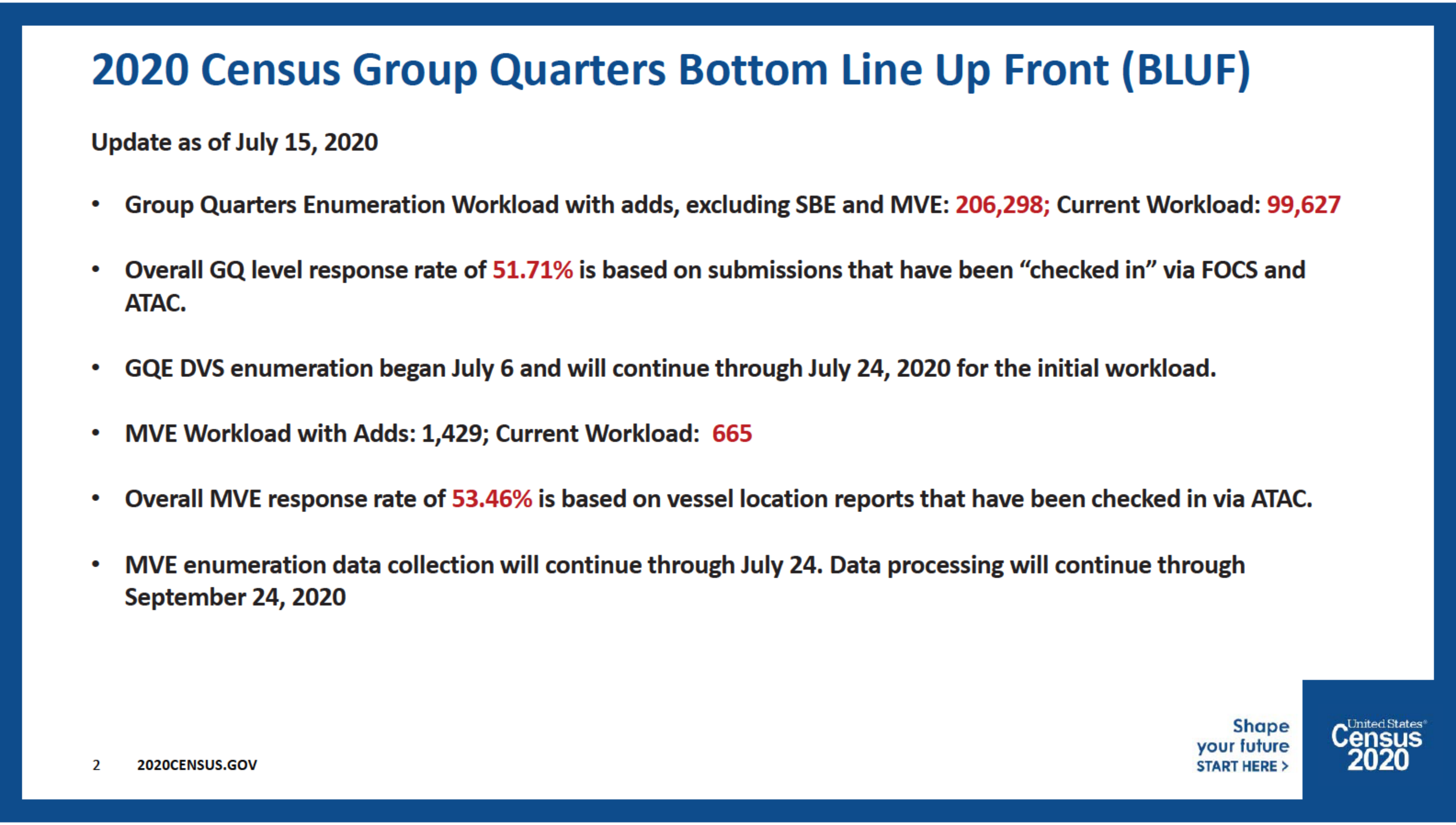
Presented by: Dora Durante and Crystal Miller

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

Shape
your future
START HERE >

United States®
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002

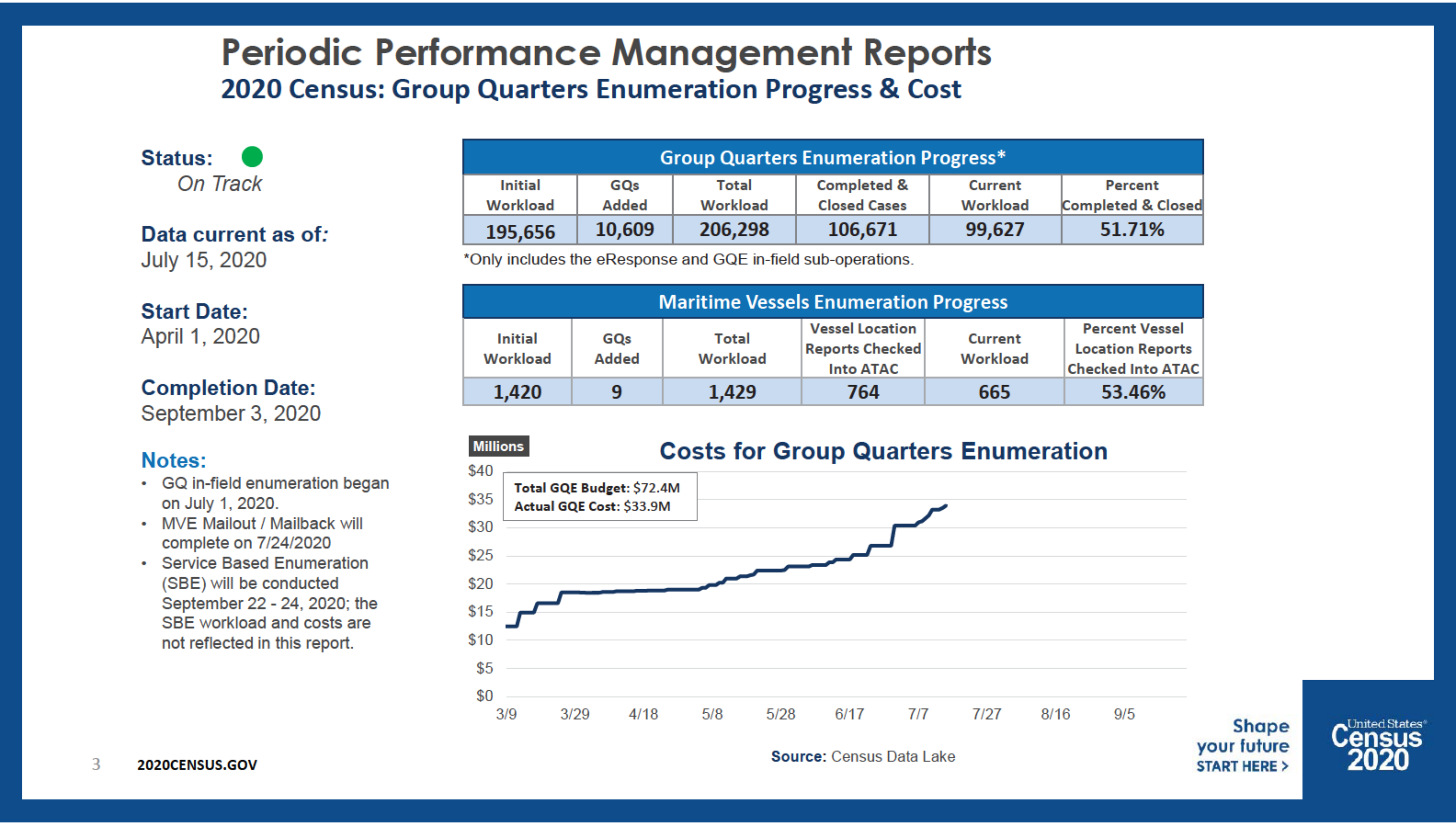


2020 Census Group Quarters Bottom Line Up Front (BLUF)

Update as of July 15, 2020

- **Group Quarters Enumeration Workload with adds, excluding SBE and MVE: 206,298; Current Workload: 99,627**
- **Overall GQ level response rate of 51.71% is based on submissions that have been "checked in" via FOCS and ATAC.**
- **GQE DVS enumeration began July 6 and will continue through July 24, 2020 for the initial workload.**
- **MVE Workload with Adds: 1,429; Current Workload: 665**
- **Overall MVE response rate of 53.46% is based on vessel location reports that have been checked in via ATAC.**
- **MVE enumeration data collection will continue through July 24. Data processing will continue through September 24, 2020**

DRB Approval Number: CBDRB-FY21-DSEP-002



Periodic Performance Management Reports

2020 Census: Group Quarters Enumeration Progress & Cost

Status: ●
On Track

Data current as of:
July 15, 2020

Start Date:
April 1, 2020

Completion Date:
September 3, 2020

Notes:

- GQ in-field enumeration began on July 1, 2020.
- MVE Mailout / Mailback will complete on 7/24/2020
- Service Based Enumeration (SBE) will be conducted September 22 - 24, 2020; the SBE workload and costs are not reflected in this report.

Group Quarters Enumeration Progress*

Initial Workload	GQs Added	Total Workload	Completed & Closed Cases	Current Workload	Percent Completed & Closed
195,656	10,609	206,298	106,671	99,627	51.71%

*Only includes the eResponse and GQE in-field sub-operations.

Maritime Vessels Enumeration Progress

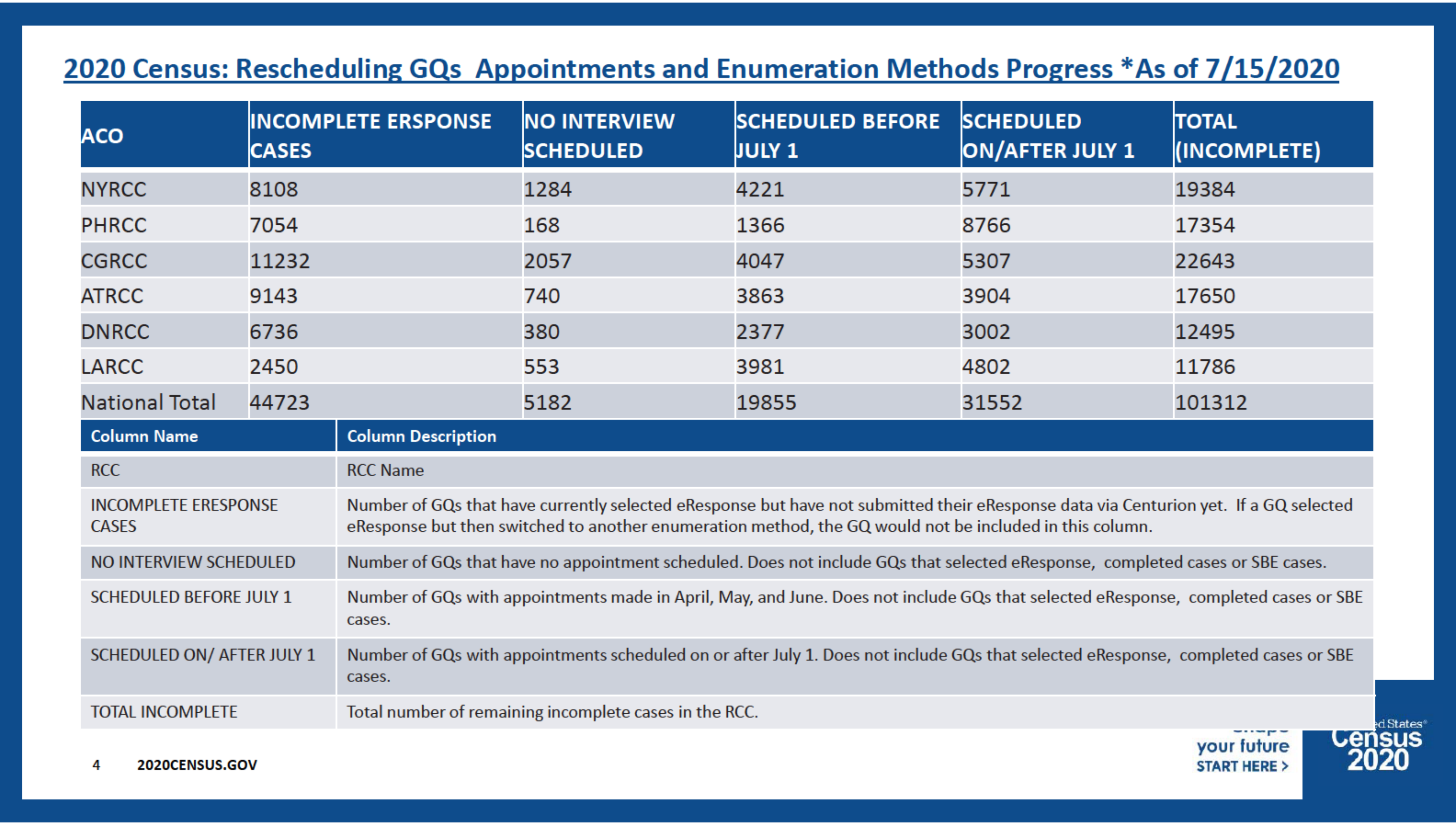
Initial Workload	GQs Added	Total Workload	Vessel Location Reports Checked Into ATAC	Current Workload	Percent Vessel Location Reports Checked Into ATAC
1,420	9	1,429	764	665	53.46%

Costs for Group Quarters Enumeration

Total GQE Budget: $72.4M
Actual GQE Cost: $33.9M

Source: Census Data Lake

DRB Approval Number: CBDRB-FY21-DSEP-002



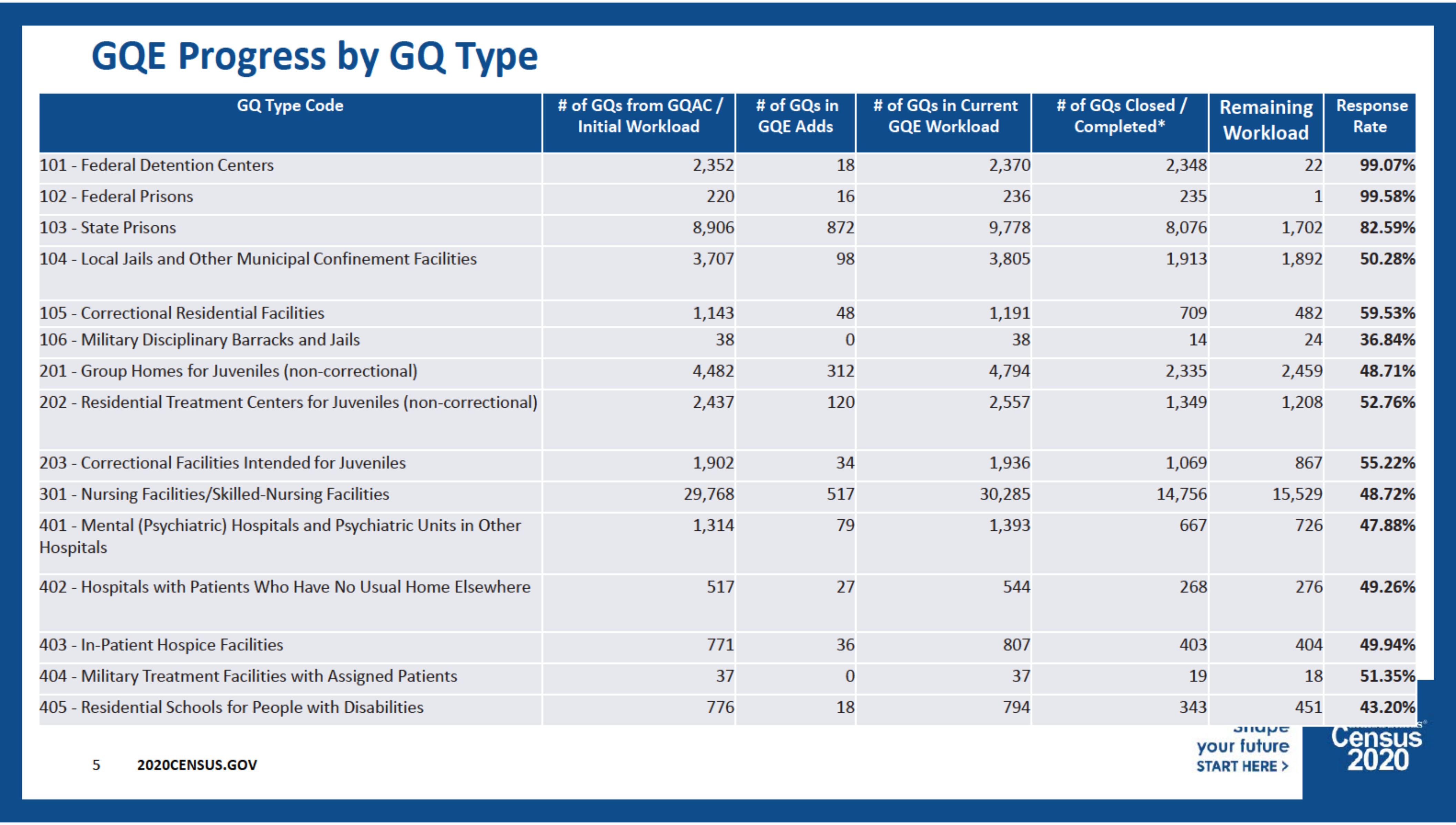
2020 Census: Rescheduling GQs Appointments and Enumeration Methods Progress *As of 7/15/2020

ACO	INCOMPLETE ERSPONSE CASES	NO INTERVIEW SCHEDULED	SCHEDULED BEFORE JULY 1	SCHEDULED ON/AFTER JULY 1	TOTAL (INCOMPLETE)
NYRCC	8108	1284	4221	5771	19384
PHRCC	7054	168	1366	8766	17354
CGRCC	11232	2057	4047	5307	22643
ATRCC	9143	740	3863	3904	17650
DNRCC	6736	380	2377	3002	12495
LARCC	2450	553	3981	4802	11786
National Total	44723	5182	19855	31552	101312

Column Name	Column Description
RCC	RCC Name
INCOMPLETE ERESPONSE CASES	Number of GQs that have currently selected eResponse but have not submitted their eResponse data via Centurion yet. If a GQ selected eResponse but then switched to another enumeration method, the GQ would not be included in this column.
NO INTERVIEW SCHEDULED	Number of GQs that have no appointment scheduled. Does not include GQs that selected eResponse, completed cases or SBE cases.
SCHEDULED BEFORE JULY 1	Number of GQs with appointments made in April, May, and June. Does not include GQs that selected eResponse, completed cases or SBE cases.
SCHEDULED ON/ AFTER JULY 1	Number of GQs with appointments scheduled on or after July 1. Does not include GQs that selected eResponse, completed cases or SBE cases.
TOTAL INCOMPLETE	Total number of remaining incomplete cases in the RCC.

your future
START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



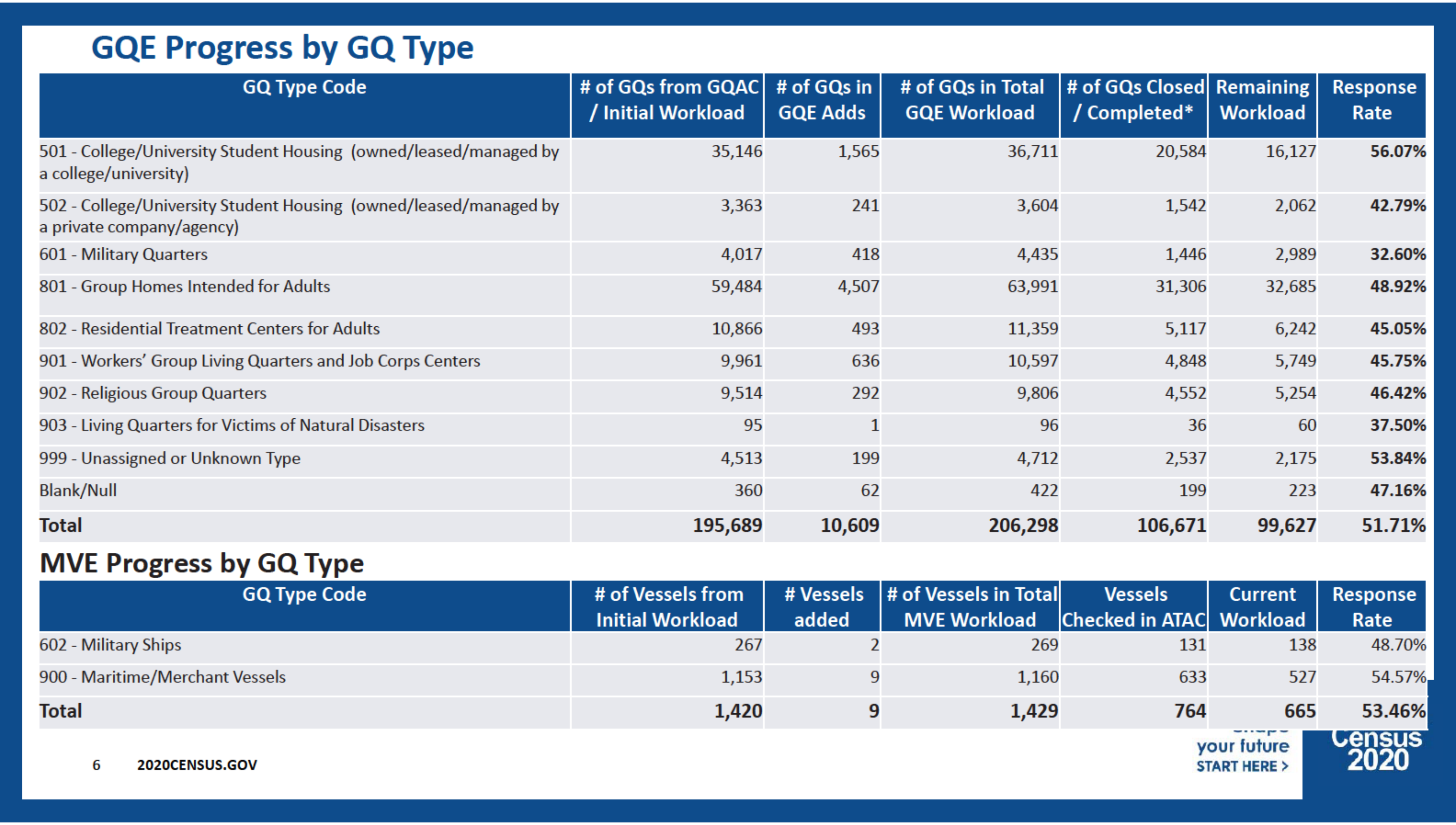
GQE Progress by GQ Type

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Current GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
101 - Federal Detention Centers	2,352	18	2,370	2,348	22	**99.07%**
102 - Federal Prisons	220	16	236	235	1	**99.58%**
103 - State Prisons	8,906	872	9,778	8,076	1,702	**82.59%**
104 - Local Jails and Other Municipal Confinement Facilities	3,707	98	3,805	1,913	1,892	**50.28%**
105 - Correctional Residential Facilities	1,143	48	1,191	709	482	**59.53%**
106 - Military Disciplinary Barracks and Jails	38	0	38	14	24	**36.84%**
201 - Group Homes for Juveniles (non-correctional)	4,482	312	4,794	2,335	2,459	**48.71%**
202 - Residential Treatment Centers for Juveniles (non-correctional)	2,437	120	2,557	1,349	1,208	**52.76%**
203 - Correctional Facilities Intended for Juveniles	1,902	34	1,936	1,069	867	**55.22%**
301 - Nursing Facilities/Skilled-Nursing Facilities	29,768	517	30,285	14,756	15,529	**48.72%**
401 - Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals	1,314	79	1,393	667	726	**47.88%**
402 - Hospitals with Patients Who Have No Usual Home Elsewhere	517	27	544	268	276	**49.26%**
403 - In-Patient Hospice Facilities	771	36	807	403	404	**49.94%**
404 - Military Treatment Facilities with Assigned Patients	37	0	37	19	18	**51.35%**
405 - Residential Schools for People with Disabilities	776	18	794	343	451	**43.20%**

Shape your future START HERE >

Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



GQE Progress by GQ Type

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Total GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
501 - College/University Student Housing (owned/leased/managed by a college/university)	35,146	1,565	36,711	20,584	16,127	**56.07%**
502 - College/University Student Housing (owned/leased/managed by a private company/agency)	3,363	241	3,604	1,542	2,062	**42.79%**
601 - Military Quarters	4,017	418	4,435	1,446	2,989	**32.60%**
801 - Group Homes Intended for Adults	59,484	4,507	63,991	31,306	32,685	**48.92%**
802 - Residential Treatment Centers for Adults	10,866	493	11,359	5,117	6,242	**45.05%**
901 - Workers' Group Living Quarters and Job Corps Centers	9,961	636	10,597	4,848	5,749	**45.75%**
902 - Religious Group Quarters	9,514	292	9,806	4,552	5,254	**46.42%**
903 - Living Quarters for Victims of Natural Disasters	95	1	96	36	60	**37.50%**
999 - Unassigned or Unknown Type	4,513	199	4,712	2,537	2,175	**53.84%**
Blank/Null	360	62	422	199	223	**47.16%**
Total	**195,689**	**10,609**	**206,298**	**106,671**	**99,627**	**51.71%**

MVE Progress by GQ Type

GQ Type Code	# of Vessels from Initial Workload	# Vessels added	# of Vessels in Total MVE Workload	Vessels Checked in ATAC	Current Workload	Response Rate
602 - Military Ships	267	2	269	131	138	48.70%
900 - Maritime/Merchant Vessels	1,153	9	1,160	633	527	54.57%
Total	**1,420**	**9**	**1,429**	**764**	**665**	**53.46%**

Issues/ Exceptions

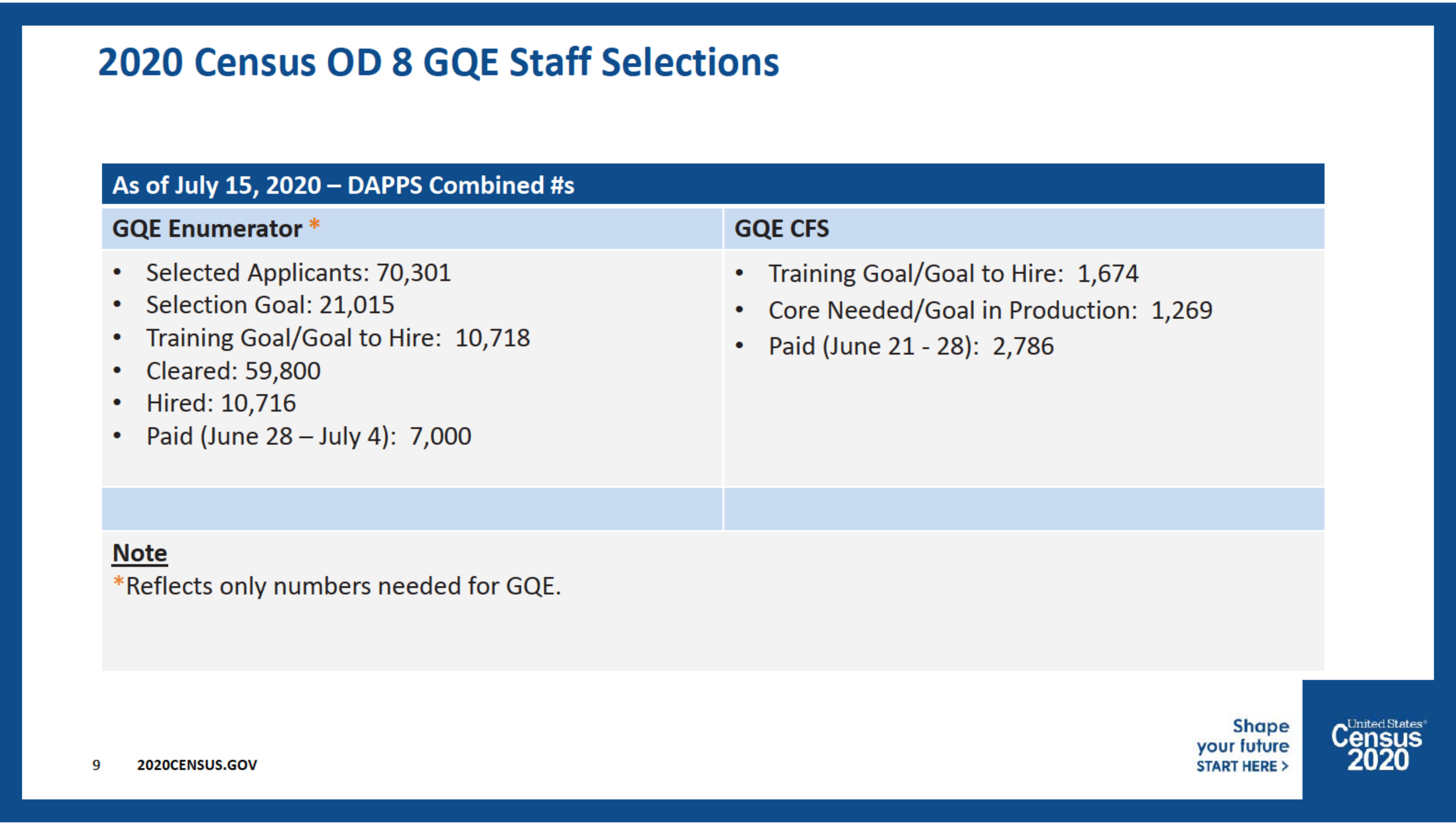
Based on guidelines given on June 24 for Restarting operations: "-- *GQE field work starts on July 1, as you well know. The decision is to proceed with GQs that we speak with via phone and who are willing to have us come out, beginning July 1. For GQs that we are not able to contact by phone, we will not proceed to the field until we start NRFU in those states (those states will be updated every Thursday at CIG, as I mentioned above). For these cases with no phone contact, the earliest field visit would be July 16."*

- **Only 6 ACOs can make personal visit follow-up on Unresolved cases or cases without appointments. All other ACOs can only visit GQs that have scheduled appointments.**
- **May need to adjust finish date for DVS data collection due to date/timing needed by The Salvation Army to put out their Directives. They requested no contact prior to July 22 to allow Directives to be delivered to all of their entities to assure cooperation and knowledge of preferred enumeration method (drop off/pick up).**

Backup

Schedules and Systems

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census OD 8 GQE Staff Selections

As of July 15, 2020 – DAPPS Combined #s	
GQE Enumerator *	**GQE CFS**
• Selected Applicants: 70,301 • Selection Goal: 21,015 • Training Goal/Goal to Hire: 10,718 • Cleared: 59,800 • Hired: 10,716 • Paid (June 28 – July 4): 7,000	• Training Goal/Goal to Hire: 1,674 • Core Needed/Goal in Production: 1,269 • Paid (June 21 - 28): 2,786

Note

*Reflects only numbers needed for GQE.

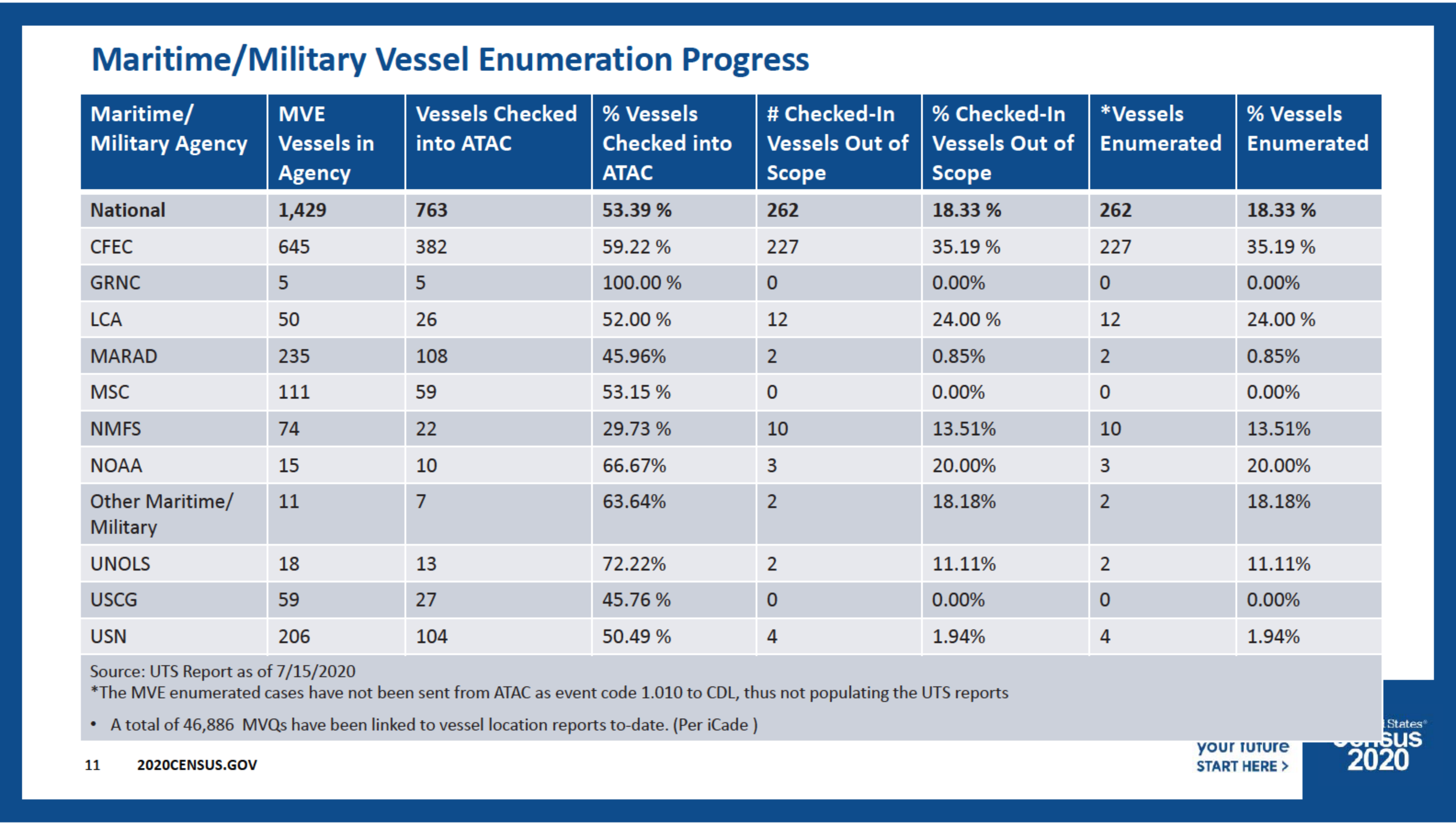
Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

OD 8 GQE Milestone Conduct & Training Dates

Conduct Activity	Proposed Start	Proposed Finish
Conduct MVE Operation	04/01/20 (A)	07/24/20 (P)
Conduct GQE eResponse Operation	04/01/20 (A)	08/07/20 (P)
Conduct GQE Operation Field	04/20/20 (A)	08/26/20 (P)
Conduct GQE In-Person Operation	07/01/20 (A)	08/26/20 (P)
Conduct GQ Reinterview Operation	07/02/20 (A)	09/03/20 (P)
Conduct DVS Enumeration Field	07/06/20 (A)	07/24/20 (P)



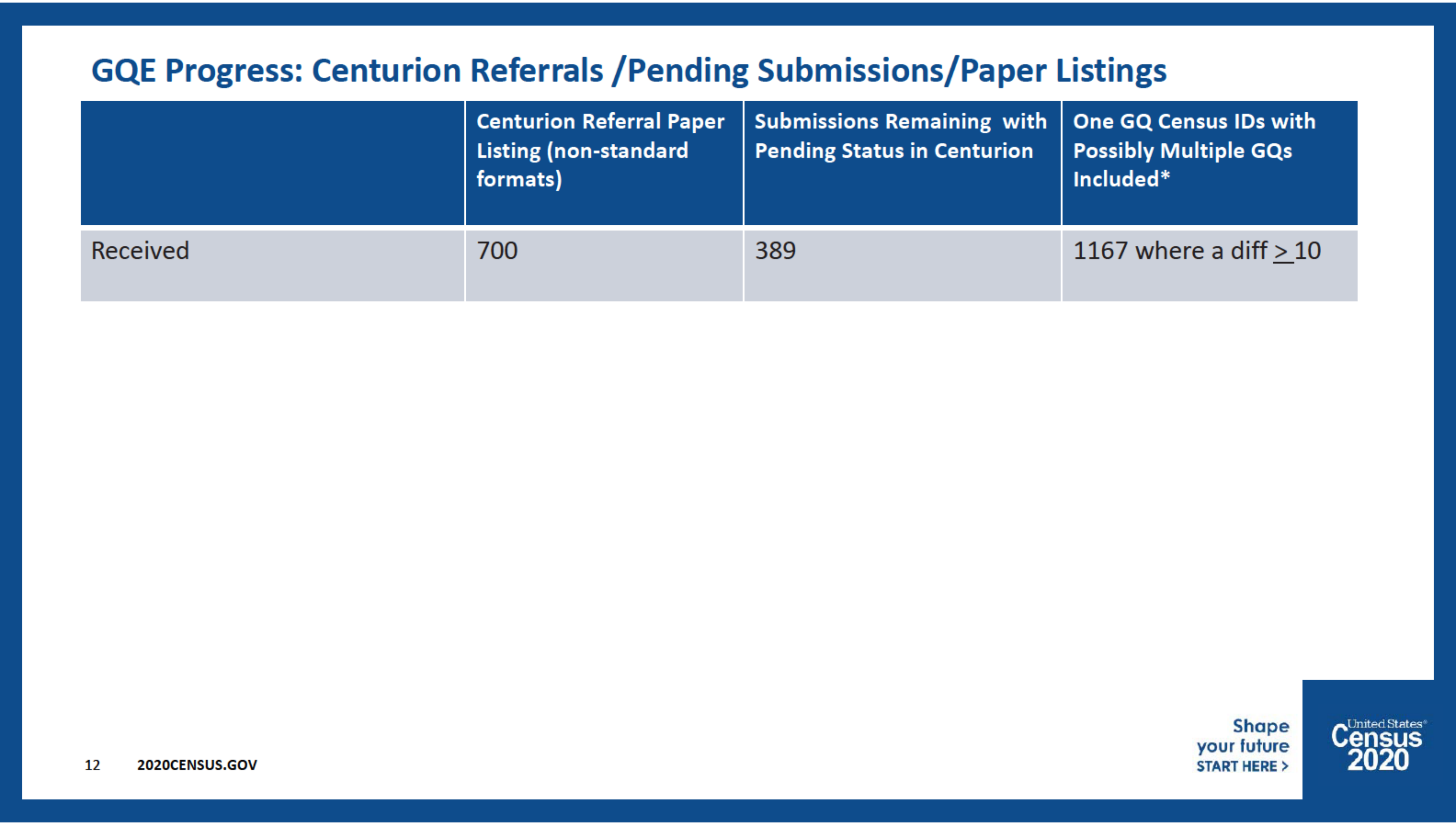
Maritime/Military Vessel Enumeration Progress

Maritime/ Military Agency	MVE Vessels in Agency	Vessels Checked into ATAC	% Vessels Checked into ATAC	# Checked-In Vessels Out of Scope	% Checked-In Vessels Out of Scope	*Vessels Enumerated	% Vessels Enumerated
National	**1,429**	**763**	**53.39 %**	**262**	**18.33 %**	**262**	**18.33 %**
CFEC	645	382	59.22 %	227	35.19 %	227	35.19 %
GRNC	5	5	100.00 %	0	0.00%	0	0.00%
LCA	50	26	52.00 %	12	24.00 %	12	24.00 %
MARAD	235	108	45.96%	2	0.85%	2	0.85%
MSC	111	59	53.15 %	0	0.00%	0	0.00%
NMFS	74	22	29.73 %	10	13.51%	10	13.51%
NOAA	15	10	66.67%	3	20.00%	3	20.00%
Other Maritime/ Military	11	7	63.64%	2	18.18%	2	18.18%
UNOLS	18	13	72.22%	2	11.11%	2	11.11%
USCG	59	27	45.76 %	0	0.00%	0	0.00%
USN	206	104	50.49 %	4	1.94%	4	1.94%

Source: UTS Report as of 7/15/2020
*The MVE enumerated cases have not been sent from ATAC as event code 1.010 to CDL, thus not populating the UTS reports

- A total of 46,886 MVQs have been linked to vessel location reports to-date. (Per iCade)

11 2020CENSUS.GOV

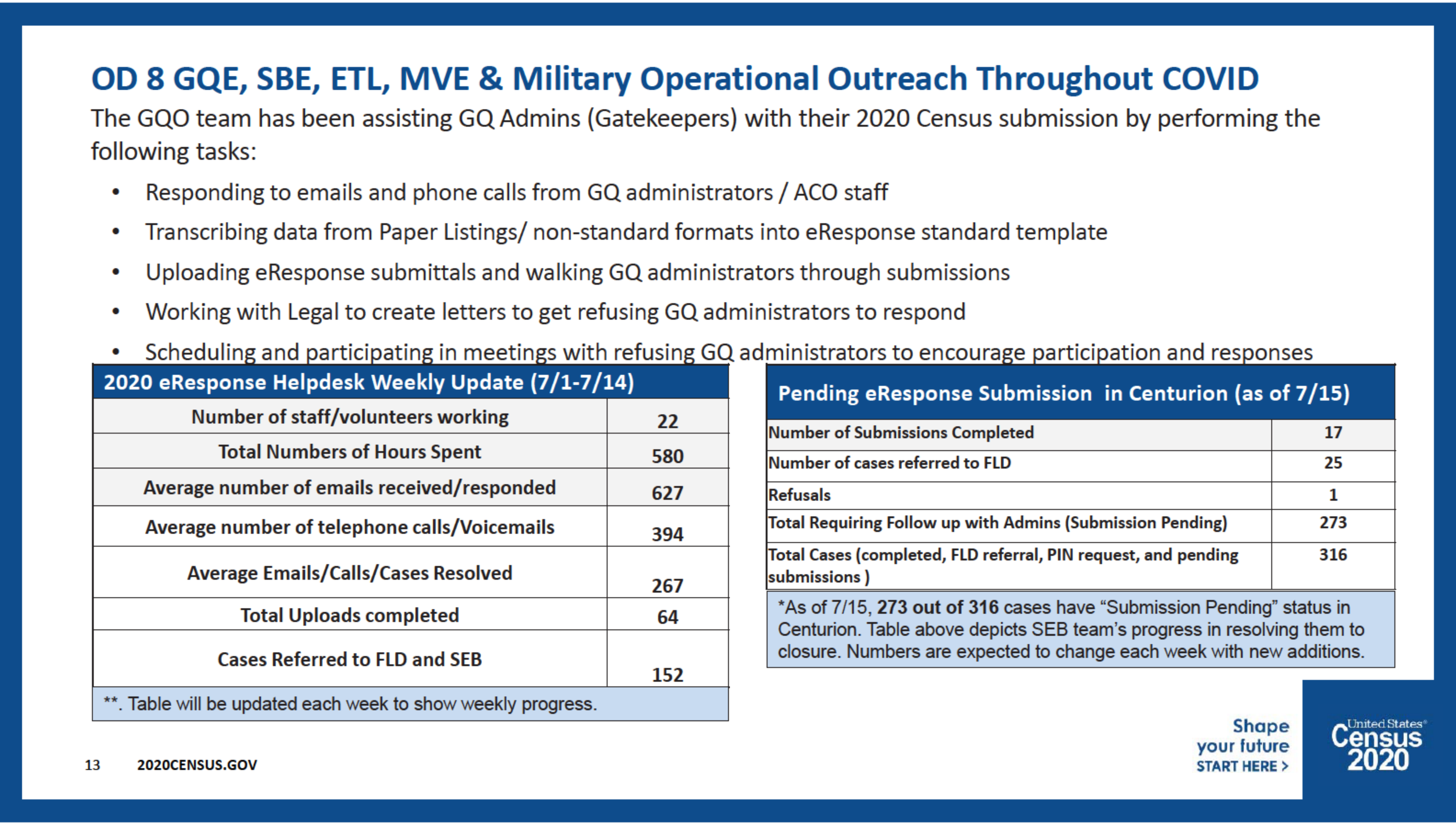


GQE Progress: Centurion Referrals /Pending Submissions/Paper Listings

	Centurion Referral Paper Listing (non-standard formats)	Submissions Remaining with Pending Status in Centurion	One GQ Census IDs with Possibly Multiple GQs Included*
Received	700	389	1167 where a diff ≥ 10

Shape
your future
START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



OD 8 GQE, SBE, ETL, MVE & Military Operational Outreach Throughout COVID

The GQO team has been assisting GQ Admins (Gatekeepers) with their 2020 Census submission by performing the following tasks:

- Responding to emails and phone calls from GQ administrators / ACO staff
- Transcribing data from Paper Listings/ non-standard formats into eResponse standard template
- Uploading eResponse submittals and walking GQ administrators through submissions
- Working with Legal to create letters to get refusing GQ administrators to respond
- Scheduling and participating in meetings with refusing GQ administrators to encourage participation and responses

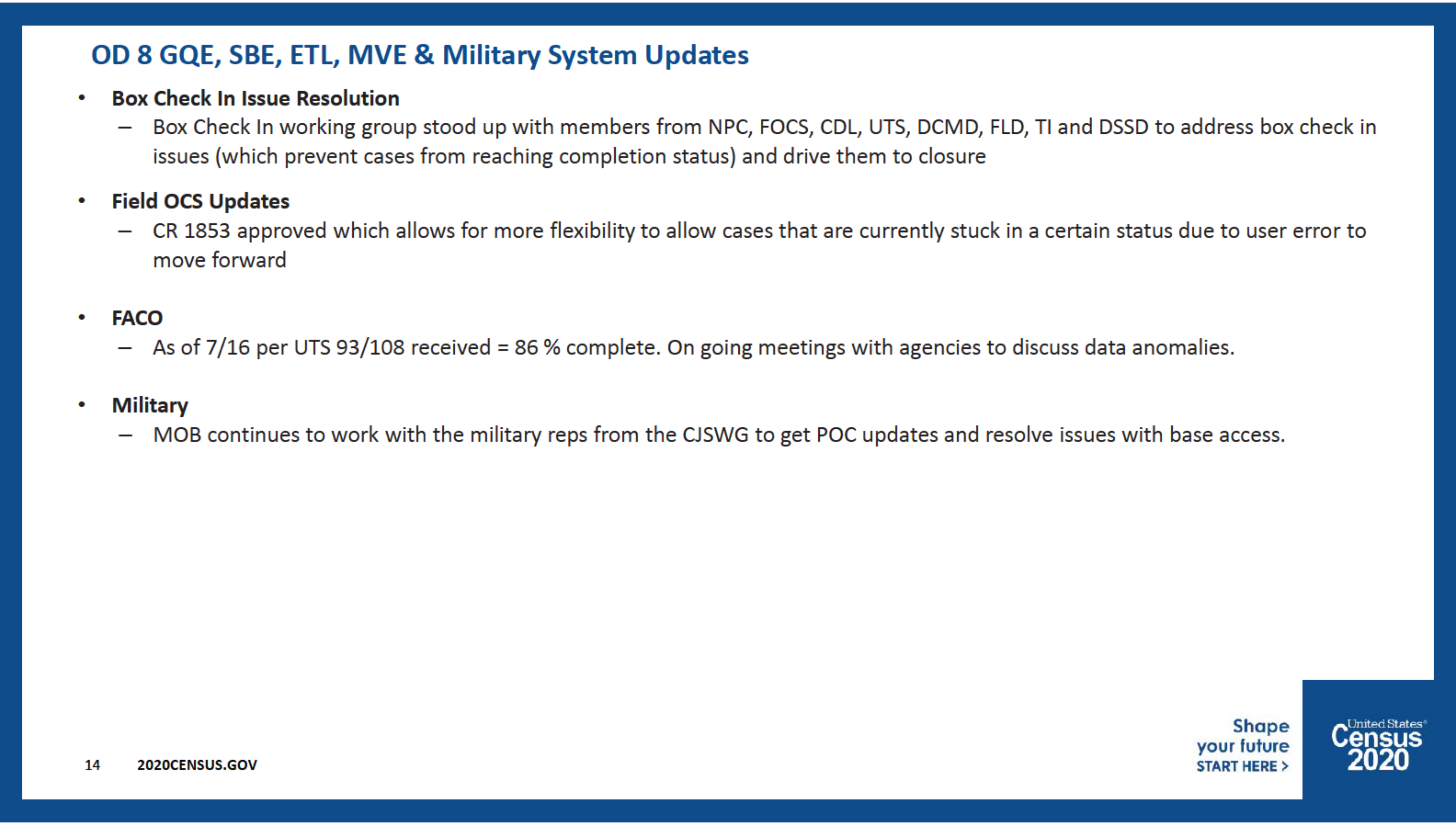
2020 eResponse Helpdesk Weekly Update (7/1-7/14)	
Number of staff/volunteers working	22
Total Numbers of Hours Spent	580
Average number of emails received/responded	627
Average number of telephone calls/Voicemails	394
Average Emails/Calls/Cases Resolved	267
Total Uploads completed	64
Cases Referred to FLD and SEB	152

**. Table will be updated each week to show weekly progress.

Pending eResponse Submission in Centurion (as of 7/15)	
Number of Submissions Completed	17
Number of cases referred to FLD	25
Refusals	1
Total Requiring Follow up with Admins (Submission Pending)	273
Total Cases (completed, FLD referral, PIN request, and pending submissions)	316

*As of 7/15, **273 out of 316** cases have "Submission Pending" status in Centurion. Table above depicts SEB team's progress in resolving them to closure. Numbers are expected to change each week with new additions.

DRB Approval Number: CBDRB-FY21-DSEP-002

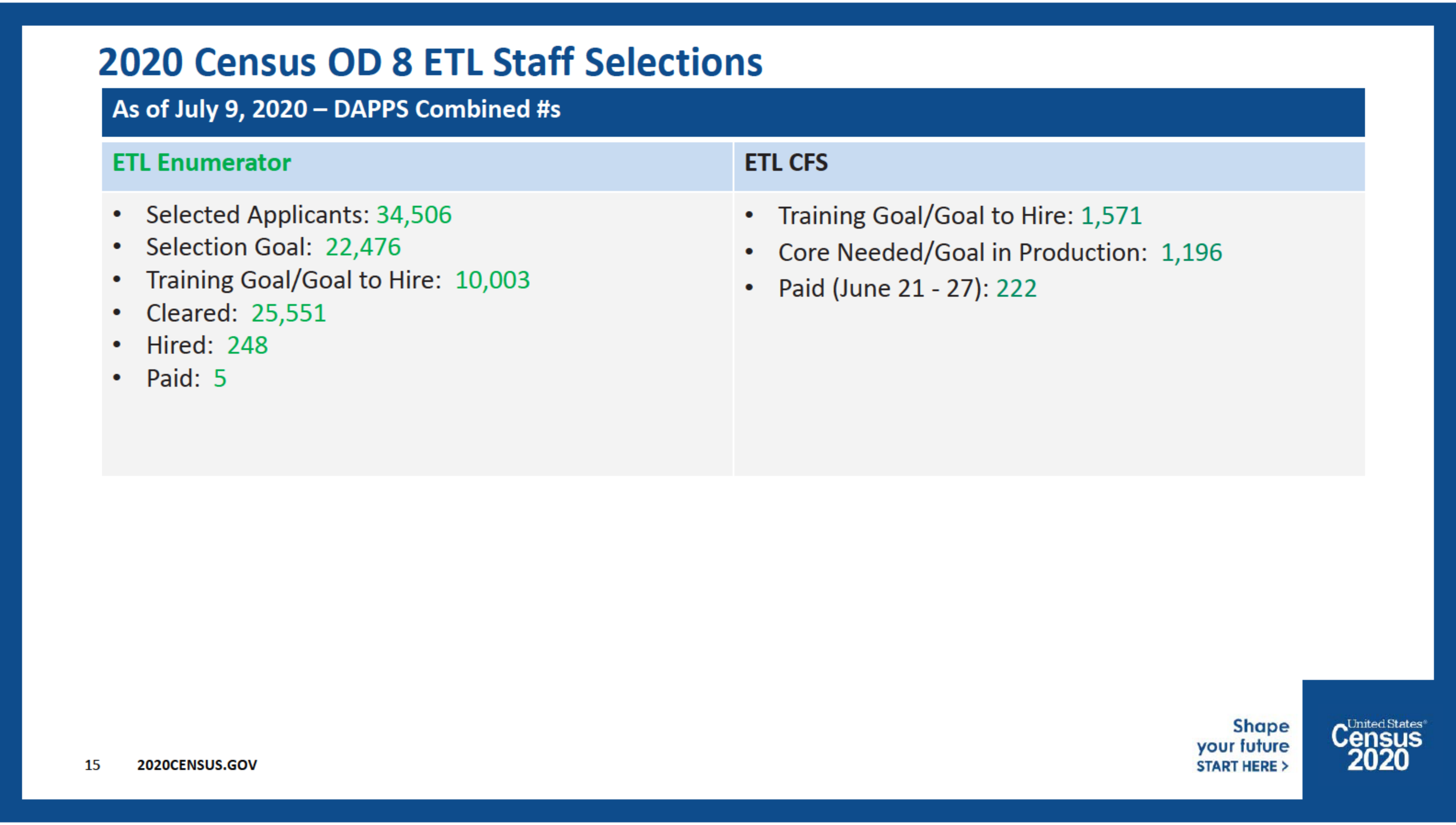


OD 8 GQE, SBE, ETL, MVE & Military System Updates

- **Box Check In Issue Resolution**
  - Box Check In working group stood up with members from NPC, FOCS, CDL, UTS, DCMD, FLD, TI and DSSD to address box check in issues (which prevent cases from reaching completion status) and drive them to closure
- **Field OCS Updates**
  - CR 1853 approved which allows for more flexibility to allow cases that are currently stuck in a certain status due to user error to move forward
- **FACO**
  - As of 7/16 per UTS 93/108 received = 86 % complete. On going meetings with agencies to discuss data anomalies.
- **Military**
  - MOB continues to work with the military reps from the CJSWG to get POC updates and resolve issues with base access.

Shape your future START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census OD 8 ETL Staff Selections

As of July 9, 2020 – DAPPS Combined #s

ETL Enumerator

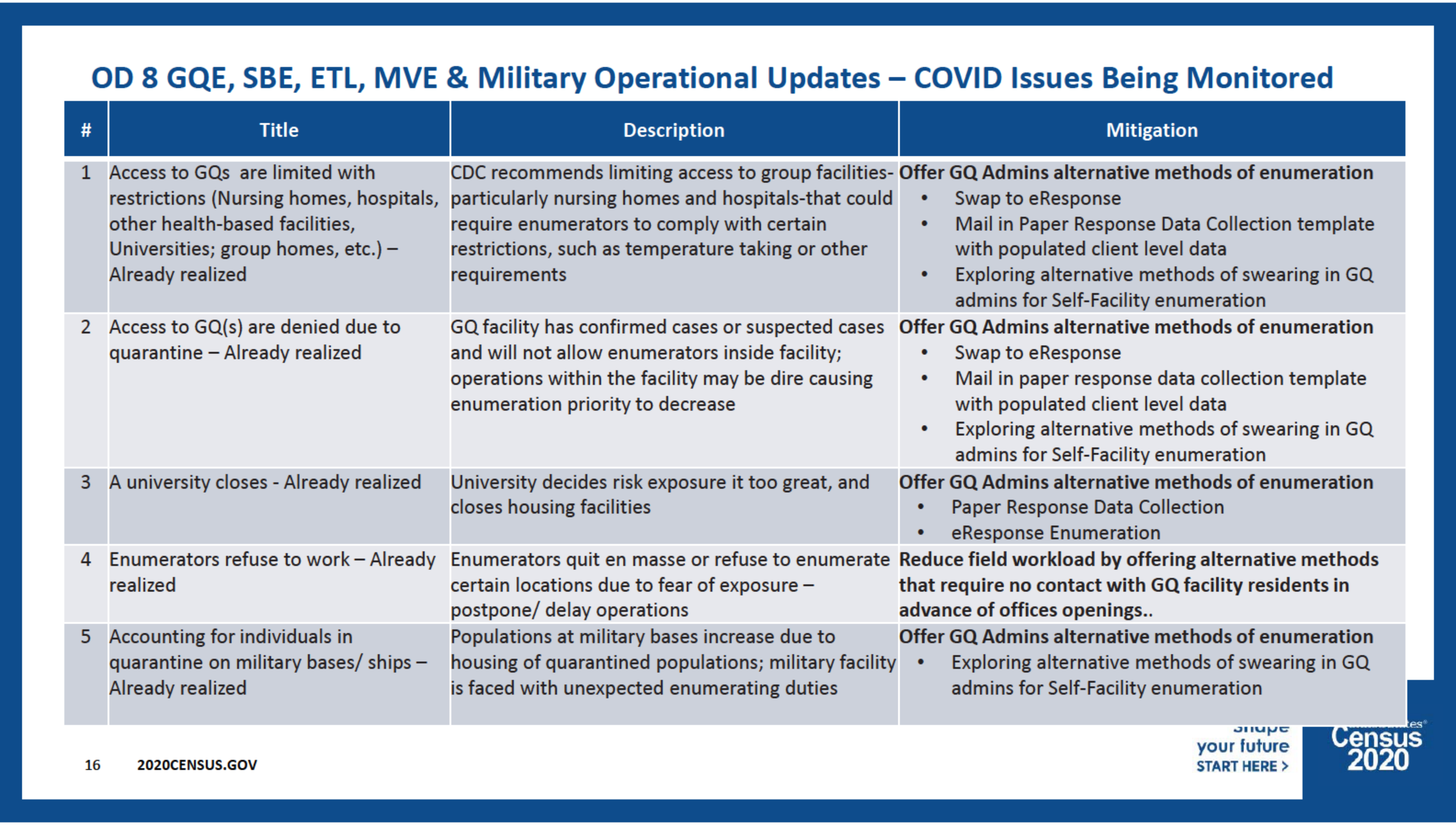
- Selected Applicants: 34,506
- Selection Goal: 22,476
- Training Goal/Goal to Hire: 10,003
- Cleared: 25,551
- Hired: 248
- Paid: 5

ETL CFS

- Training Goal/Goal to Hire: 1,571
- Core Needed/Goal in Production: 1,196
- Paid (June 21 - 27): 222

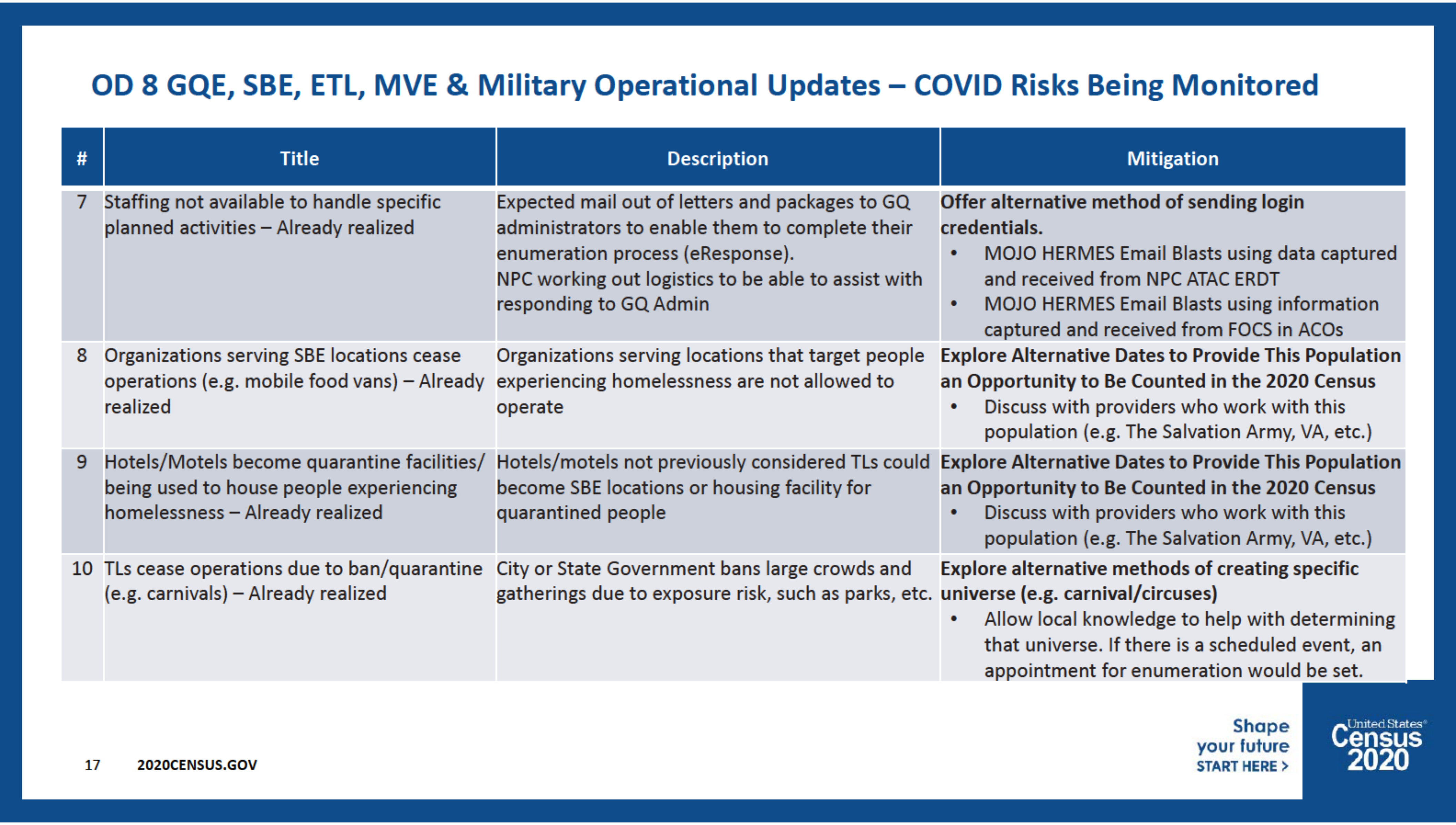
Shape
your future
START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Issues Being Monitored

#	Title	Description	Mitigation
1	Access to GQs are limited with restrictions (Nursing homes, hospitals, other health-based facilities, Universities; group homes, etc.) – Already realized	CDC recommends limiting access to group facilities-particularly nursing homes and hospitals-that could require enumerators to comply with certain restrictions, such as temperature taking or other requirements	**Offer GQ Admins alternative methods of enumeration** • Swap to eResponse • Mail in Paper Response Data Collection template with populated client level data • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration
2	Access to GQ(s) are denied due to quarantine – Already realized	GQ facility has confirmed cases or suspected cases and will not allow enumerators inside facility; operations within the facility may be dire causing enumeration priority to decrease	**Offer GQ Admins alternative methods of enumeration** • Swap to eResponse • Mail in paper response data collection template with populated client level data • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration
3	A university closes - Already realized	University decides risk exposure it too great, and closes housing facilities	**Offer GQ Admins alternative methods of enumeration** • Paper Response Data Collection • eResponse Enumeration
4	Enumerators refuse to work – Already realized	Enumerators quit en masse or refuse to enumerate certain locations due to fear of exposure – postpone/ delay operations	**Reduce field workload by offering alternative methods that require no contact with GQ facility residents in advance of offices openings..**
5	Accounting for individuals in quarantine on military bases/ ships – Already realized	Populations at military bases increase due to housing of quarantined populations; military facility is faced with unexpected enumerating duties	**Offer GQ Admins alternative methods of enumeration** • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration



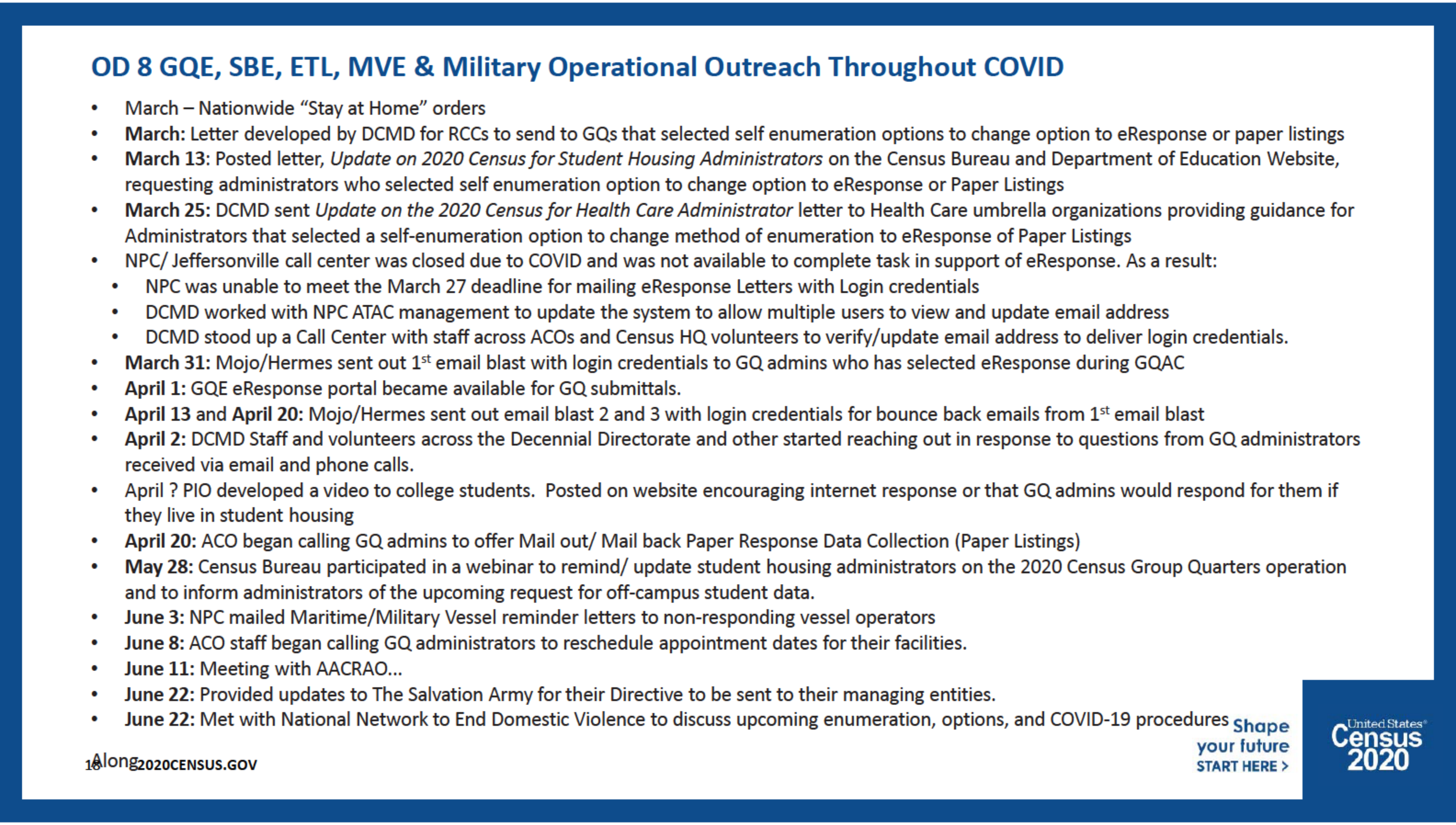
OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Risks Being Monitored

#	Title	Description	Mitigation
7	Staffing not available to handle specific planned activities – Already realized	Expected mail out of letters and packages to GQ administrators to enable them to complete their enumeration process (eResponse). NPC working out logistics to be able to assist with responding to GQ Admin	**Offer alternative method of sending login credentials.** • MOJO HERMES Email Blasts using data captured and received from NPC ATAC ERDT • MOJO HERMES Email Blasts using information captured and received from FOCS in ACOs
8	Organizations serving SBE locations cease operations (e.g. mobile food vans) – Already realized	Organizations serving locations that target people experiencing homelessness are not allowed to operate	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
9	Hotels/Motels become quarantine facilities/ being used to house people experiencing homelessness – Already realized	Hotels/motels not previously considered TLs could become SBE locations or housing facility for quarantined people	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
10	TLs cease operations due to ban/quarantine (e.g. carnivals) – Already realized	City or State Government bans large crowds and gatherings due to exposure risk, such as parks, etc.	**Explore alternative methods of creating specific universe (e.g. carnival/circuses)** • Allow local knowledge to help with determining that universe. If there is a scheduled event, an appointment for enumeration would be set.

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



OD 8 GQE, SBE, ETL, MVE & Military Operational Outreach Throughout COVID

- March – Nationwide "Stay at Home" orders
- **March:** Letter developed by DCMD for RCCs to send to GQs that selected self enumeration options to change option to eResponse or paper listings
- **March 13:** Posted letter, *Update on 2020 Census for Student Housing Administrators* on the Census Bureau and Department of Education Website, requesting administrators who selected self enumeration option to change option to eResponse or Paper Listings
- **March 25:** DCMD sent *Update on the 2020 Census for Health Care Administrator* letter to Health Care umbrella organizations providing guidance for Administrators that selected a self-enumeration option to change method of enumeration to eResponse of Paper Listings
- NPC/ Jeffersonville call center was closed due to COVID and was not available to complete task in support of eResponse. As a result:
  - NPC was unable to meet the March 27 deadline for mailing eResponse Letters with Login credentials
  - DCMD worked with NPC ATAC management to update the system to allow multiple users to view and update email address
  - DCMD stood up a Call Center with staff across ACOs and Census HQ volunteers to verify/update email address to deliver login credentials.
- **March 31:** Mojo/Hermes sent out 1st email blast with login credentials to GQ admins who has selected eResponse during GQAC
- **April 1:** GQE eResponse portal became available for GQ submittals.
- **April 13** and **April 20:** Mojo/Hermes sent out email blast 2 and 3 with login credentials for bounce back emails from 1st email blast
- **April 2:** DCMD Staff and volunteers across the Decennial Directorate and other started reaching out in response to questions from GQ administrators received via email and phone calls.
- April ? PIO developed a video to college students. Posted on website encouraging internet response or that GQ admins would respond for them if they live in student housing
- **April 20:** ACO began calling GQ admins to offer Mail out/ Mail back Paper Response Data Collection (Paper Listings)
- **May 28:** Census Bureau participated in a webinar to remind/ update student housing administrators on the 2020 Census Group Quarters operation and to inform administrators of the upcoming request for off-campus student data.
- **June 3:** NPC mailed Maritime/Military Vessel reminder letters to non-responding vessel operators
- **June 8:** ACO staff began calling GQ administrators to reschedule appointment dates for their facilities.
- **June 11:** Meeting with AACRAO…
- **June 22:** Provided updates to The Salvation Army for their Directive to be sent to their managing entities.
- **June 22:** Met with National Network to End Domestic Violence to discuss upcoming enumeration, options, and COVID-19 procedures

18 Along 2020CENSUS.GOV

Shape your future START HERE >

United States Census 2020

OD 8 GQE, SBE, ETL, MVE & Military Operational Updates

2020census.gov website

Conducting the Count: https://2020census.gov/en/conducting-the-count.html

Counting People in Group Living Arrangements: https://2020census.gov/en/conducting-the-count/gq.html

Group Quarters Enumeration: https://2020census.gov/en/conducting-the-count/gq/gqe.html

Service-Based Enumeration: https://2020census.gov/en/conducting-the-count/gq/sbe.html

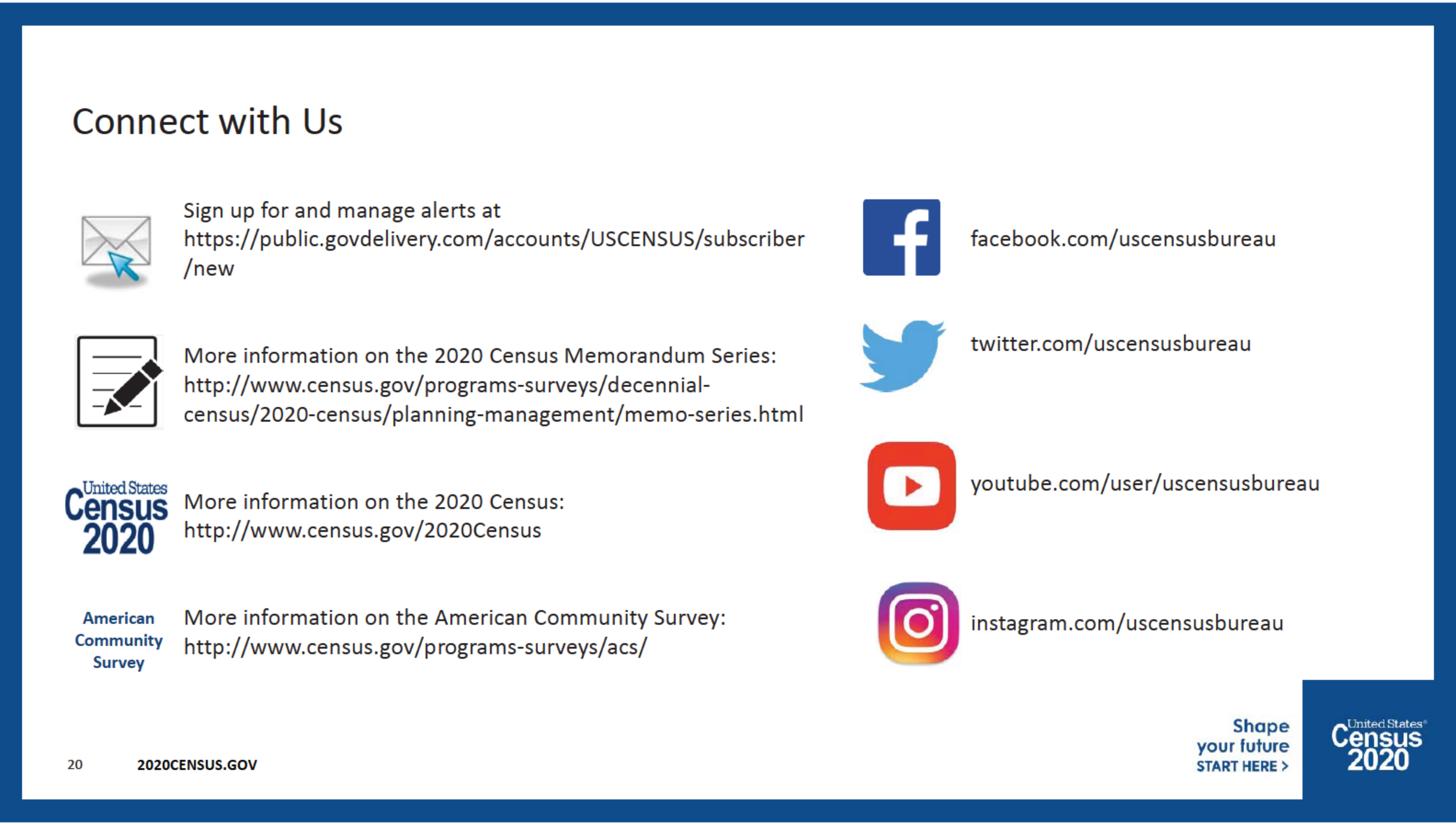
Group Quarters Advance Contact: https://2020census.gov/en/conducting-the-count/gq/gqac.html

eResponse: https://2020census.gov/en/conducting-the-count/gq/eresponse.html

Maritime and Military Vessel Enumeration: https://2020census.gov/en/conducting-the-count/gq/mve.html

Department of Education Student Privacy Policy Office: https://studentprivacy.ed.gov/faq/colleges-and-2020-census

DRB Approval Number: CBDRB-FY21-DSEP-002



Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

More information on the 2020 Census: http://www.census.gov/2020Census

American Community Survey

More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

instagram.com/uscensusbureau

20 2020CENSUS.GOV

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Operational Delivery 8: Transitory Locations

Focus: Transitory Location Advance Contact

Thursday: July 16, 2020

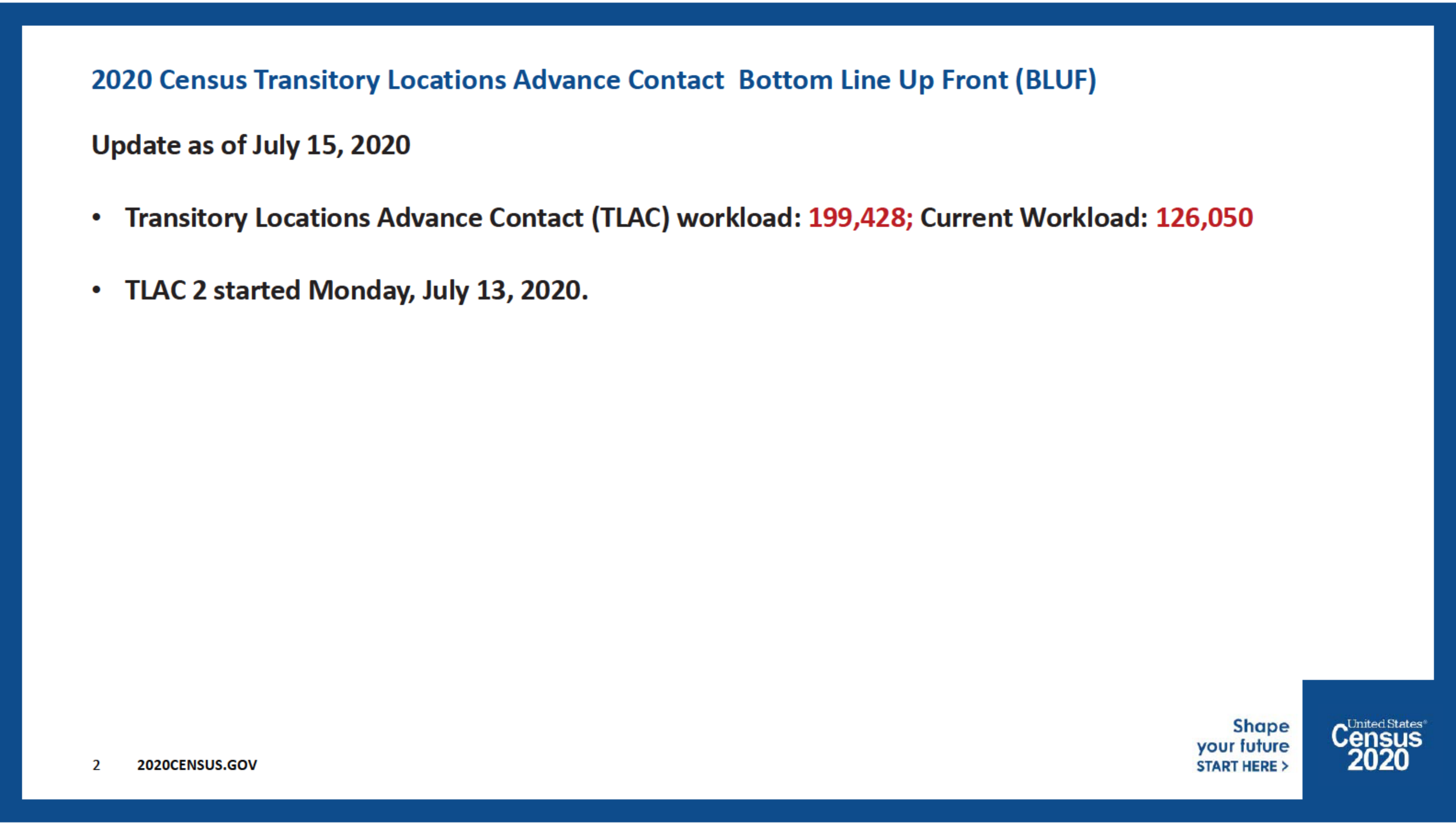
Presented by: Dora Durante, Brian Zamperini, and Crystal Miller

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

Shape
your future
START HERE >

United States®
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census Transitory Locations Advance Contact Bottom Line Up Front (BLUF)

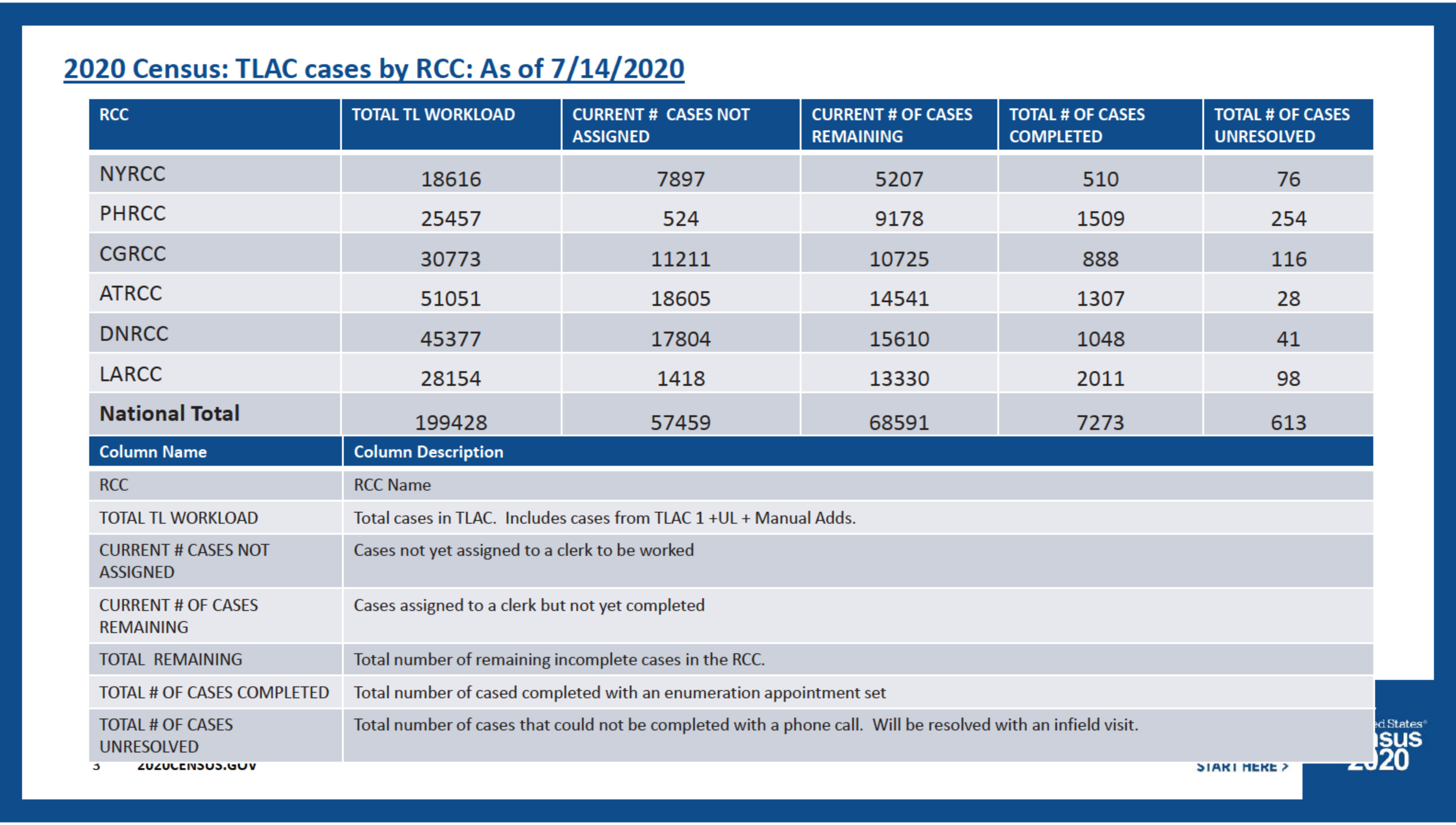
Update as of July 15, 2020

- **Transitory Locations Advance Contact (TLAC) workload: 199,428; Current Workload: 126,050**
- **TLAC 2 started Monday, July 13, 2020.**

Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census: TLAC cases by RCC: As of 7/14/2020

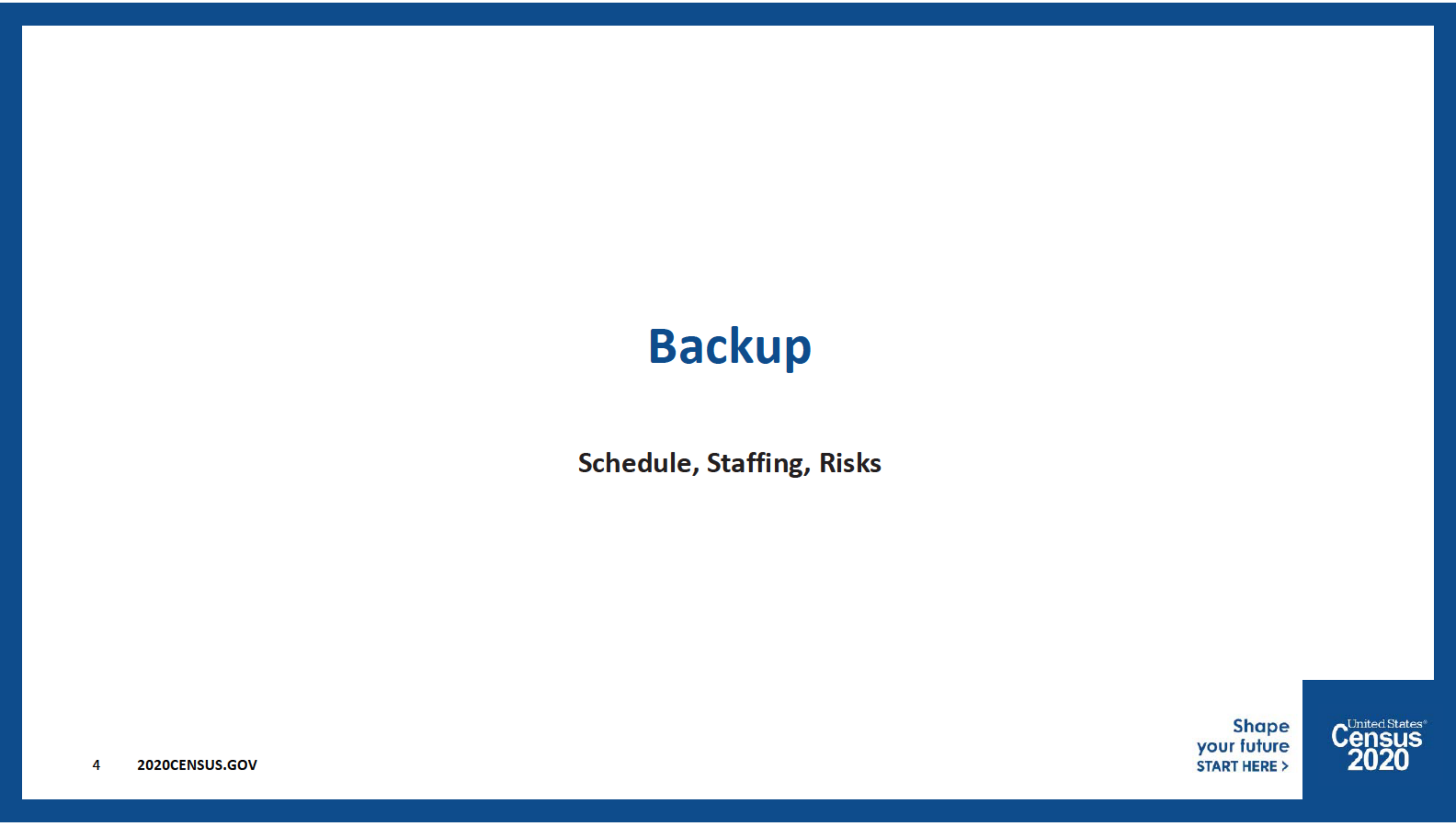
RCC	TOTAL TL WORKLOAD	CURRENT # CASES NOT ASSIGNED	CURRENT # OF CASES REMAINING	TOTAL # OF CASES COMPLETED	TOTAL # OF CASES UNRESOLVED
NYRCC	18616	7897	5207	510	76
PHRCC	25457	524	9178	1509	254
CGRCC	30773	11211	10725	888	116
ATRCC	51051	18605	14541	1307	28
DNRCC	45377	17804	15610	1048	41
LARCC	28154	1418	13330	2011	98
National Total	199428	57459	68591	7273	613

Column Name	Column Description
RCC	RCC Name
TOTAL TL WORKLOAD	Total cases in TLAC. Includes cases from TLAC 1 +UL + Manual Adds.
CURRENT # CASES NOT ASSIGNED	Cases not yet assigned to a clerk to be worked
CURRENT # OF CASES REMAINING	Cases assigned to a clerk but not yet completed
TOTAL REMAINING	Total number of remaining incomplete cases in the RCC.
TOTAL # OF CASES COMPLETED	Total number of cased completed with an enumeration appointment set
TOTAL # OF CASES UNRESOLVED	Total number of cases that could not be completed with a phone call. Will be resolved with an infield visit.

3 2020CENSUS.GOV

START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002

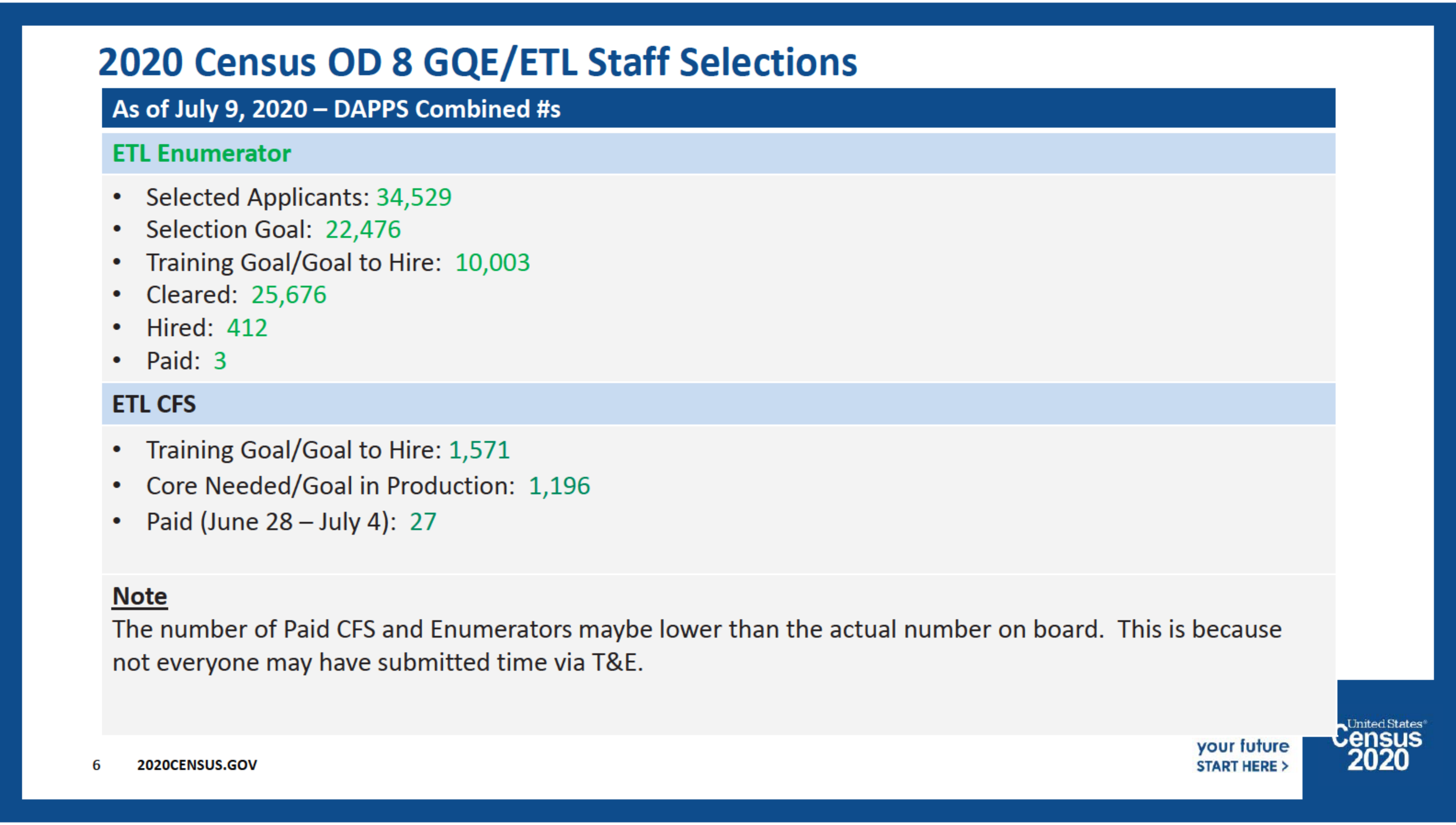


Backup

Schedule, Staffing, Risks

OD 8 GQE Milestone Conduct & Training Dates

TLAC Training Activity	Proposed Start	Proposed Finish
Conduct TLAC Clerk Refresher Training	07/09/20 (A)	7/10/20 (A)
Conduct TLAC Operation	07/13/20 (A)	08/07/20 (P)
Conduct TLAC CFS Refresher Training	07/20/20 (P)	07/21/20 (P)



2020 Census OD 8 GQE/ETL Staff Selections

As of July 9, 2020 – DAPPS Combined #s

ETL Enumerator

- Selected Applicants: 34,529
- Selection Goal: 22,476
- Training Goal/Goal to Hire: 10,003
- Cleared: 25,676
- Hired: 412
- Paid: 3

ETL CFS

- Training Goal/Goal to Hire: 1,571
- Core Needed/Goal in Production: 1,196
- Paid (June 28 – July 4): 27

Note

The number of Paid CFS and Enumerators maybe lower than the actual number on board. This is because not everyone may have submitted time via T&E.

DRB Approval Number: CBDRB-FY21-DSEP-002

OD 8 ETL Operational Updates – COVID Risks Being Monitored

#	Title	Description	Mitigation
1	Hotels/Motels become quarantine facilities/ being used to house people experiencing homelessness – Already realized	Hotels/motels not previously considered TLs could become SBE locations or housing facility for quarantined people	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
2	TLs cease operations due to ban/quarantine (e.g. carnivals) – Already realized	City or State Government bans large crowds and gatherings due to exposure risk, such as parks, etc.	**Explore alternative methods of creating specific universe (e.g. carnival/circuses)** • Allow local knowledge to help with determining that universe. If there is a scheduled event, an appointment for enumeration would be set.

Shape
your future
START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002

Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

More information on the 2020 Census: http://www.census.gov/2020Census

More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

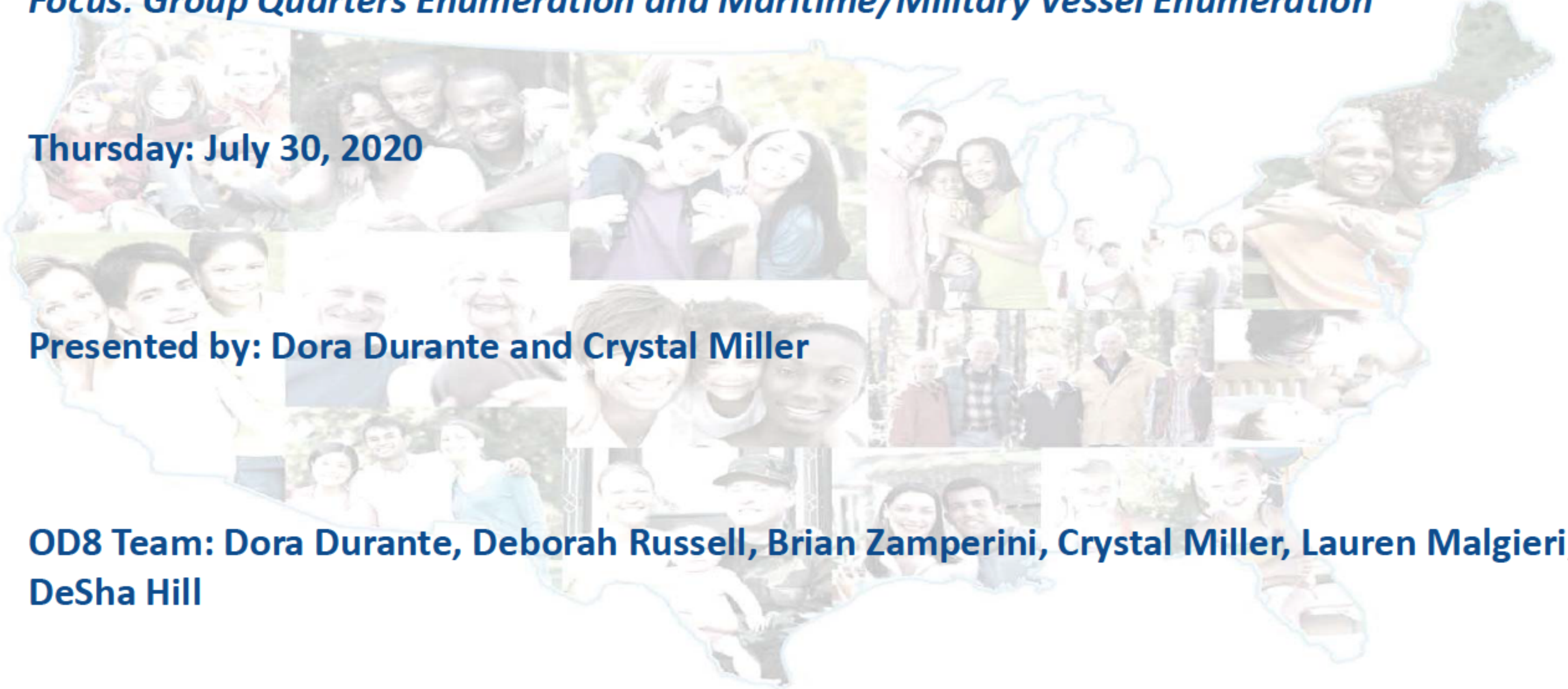
instagram.com/uscensusbureau

Shape your future START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Operational Delivery 8: Group Quarters Update

Focus: Group Quarters Enumeration and Maritime/Military Vessel Enumeration



Thursday: July 30, 2020

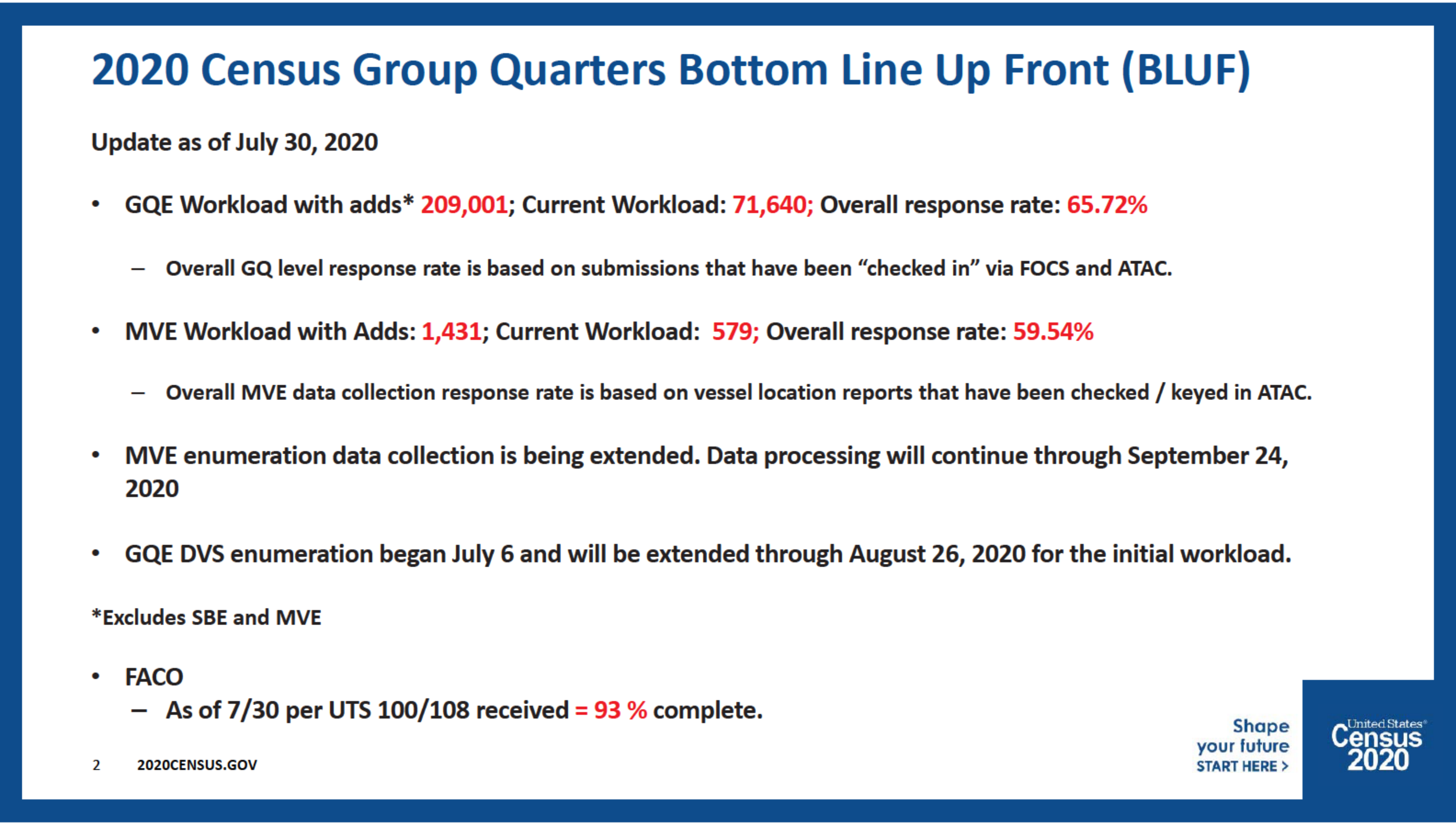
Presented by: Dora Durante and Crystal Miller

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

Shape
your future
START HERE >

United States®
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census Group Quarters Bottom Line Up Front (BLUF)

Update as of July 30, 2020

- **GQE Workload with adds* 209,001; Current Workload: 71,640; Overall response rate: 65.72%**
  - **Overall GQ level response rate is based on submissions that have been "checked in" via FOCS and ATAC.**
- **MVE Workload with Adds: 1,431; Current Workload: 579; Overall response rate: 59.54%**
  - **Overall MVE data collection response rate is based on vessel location reports that have been checked / keyed in ATAC.**
- **MVE enumeration data collection is being extended. Data processing will continue through September 24, 2020**
- **GQE DVS enumeration began July 6 and will be extended through August 26, 2020 for the initial workload.**

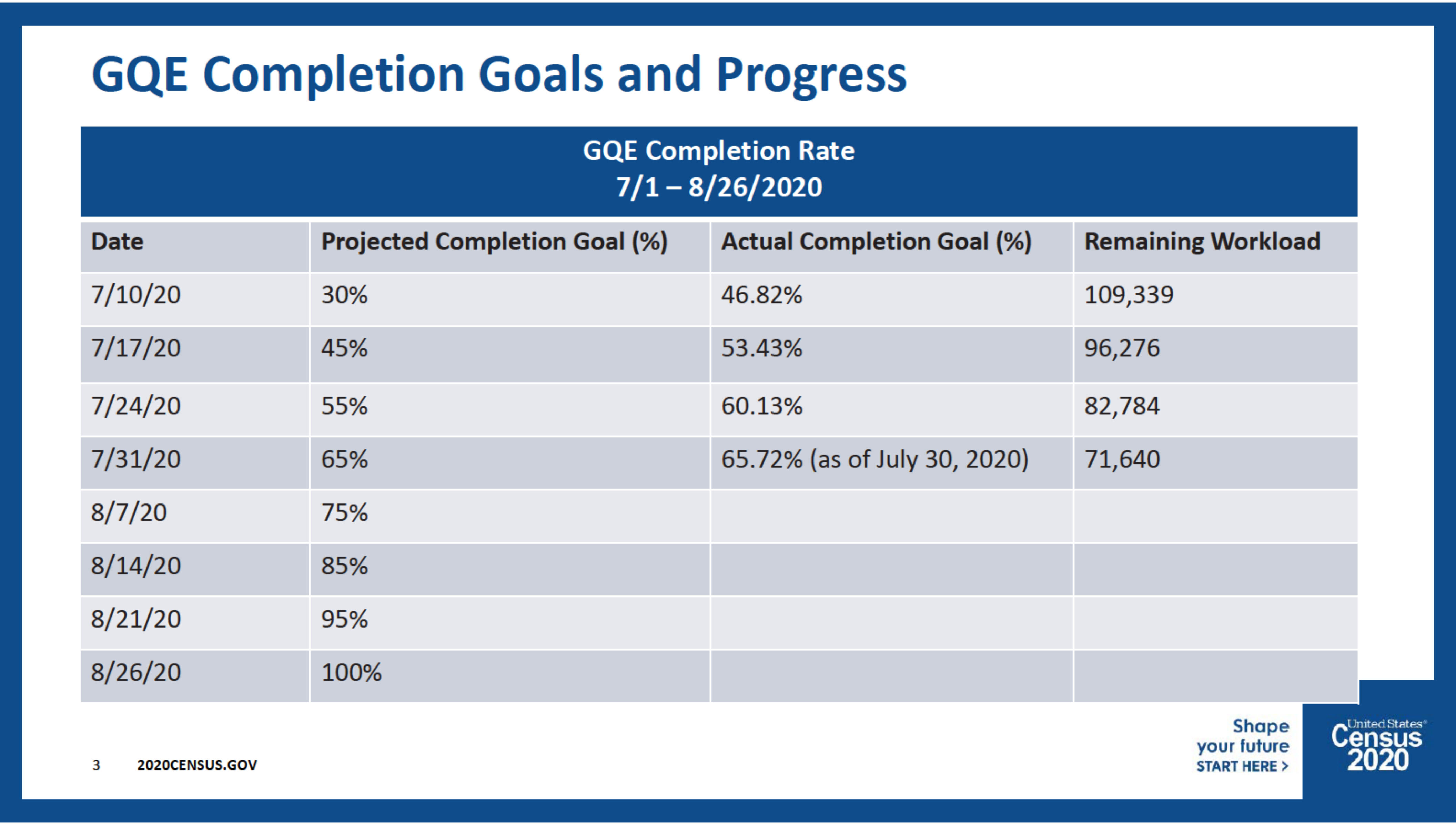
***Excludes SBE and MVE**

- **FACO**
  - **As of 7/30 per UTS 100/108 received = 93 % complete.**

Shape
your future
START HERE >

United States®
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002



GQE Completion Goals and Progress

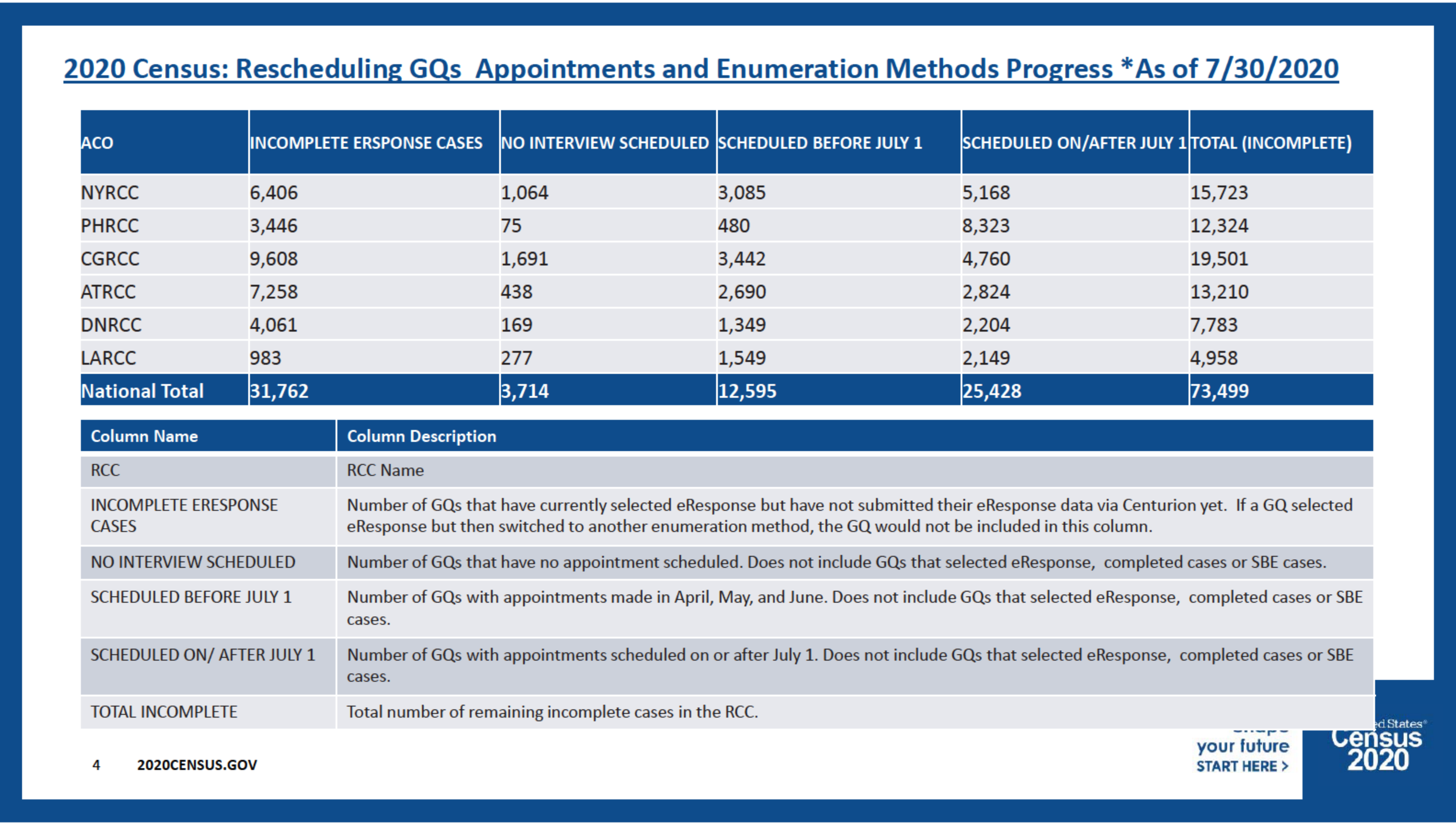
GQE Completion Rate 7/1 – 8/26/2020			
Date	**Projected Completion Goal (%)**	**Actual Completion Goal (%)**	**Remaining Workload**
7/10/20	30%	46.82%	109,339
7/17/20	45%	53.43%	96,276
7/24/20	55%	60.13%	82,784
7/31/20	65%	65.72% (as of July 30, 2020)	71,640
8/7/20	75%		
8/14/20	85%		
8/21/20	95%		
8/26/20	100%		

3 2020CENSUS.GOV

Shape your future START HERE >

United States® Census 2020

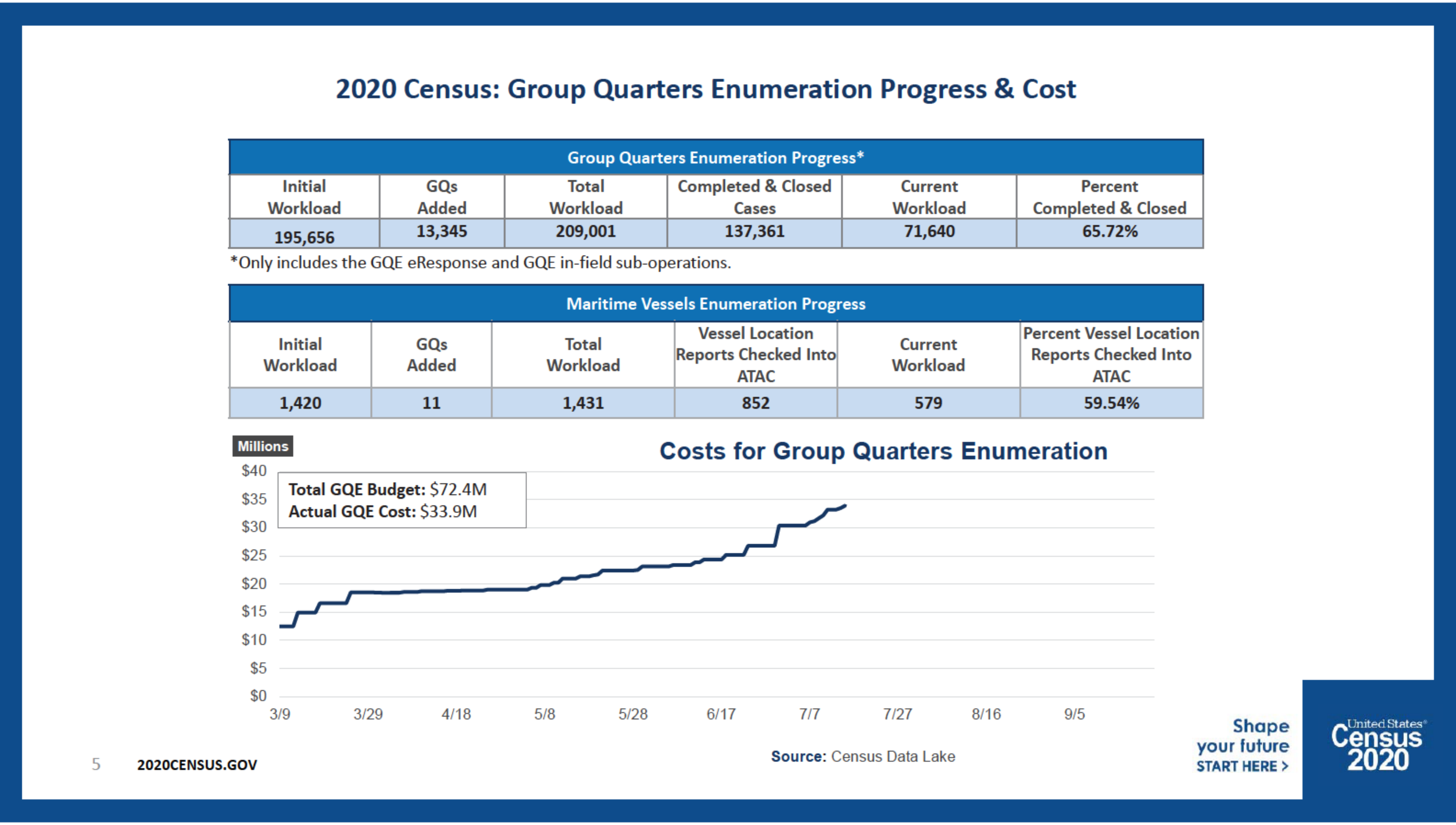
DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census: Rescheduling GQs Appointments and Enumeration Methods Progress *As of 7/30/2020

ACO	INCOMPLETE ERSPONSE CASES	NO INTERVIEW SCHEDULED	SCHEDULED BEFORE JULY 1	SCHEDULED ON/AFTER JULY 1	TOTAL (INCOMPLETE)
NYRCC	6,406	1,064	3,085	5,168	15,723
PHRCC	3,446	75	480	8,323	12,324
CGRCC	9,608	1,691	3,442	4,760	19,501
ATRCC	7,258	438	2,690	2,824	13,210
DNRCC	4,061	169	1,349	2,204	7,783
LARCC	983	277	1,549	2,149	4,958
National Total	31,762	3,714	12,595	25,428	73,499

Column Name	Column Description
RCC	RCC Name
INCOMPLETE ERESPONSE CASES	Number of GQs that have currently selected eResponse but have not submitted their eResponse data via Centurion yet. If a GQ selected eResponse but then switched to another enumeration method, the GQ would not be included in this column.
NO INTERVIEW SCHEDULED	Number of GQs that have no appointment scheduled. Does not include GQs that selected eResponse, completed cases or SBE cases.
SCHEDULED BEFORE JULY 1	Number of GQs with appointments made in April, May, and June. Does not include GQs that selected eResponse, completed cases or SBE cases.
SCHEDULED ON/ AFTER JULY 1	Number of GQs with appointments scheduled on or after July 1. Does not include GQs that selected eResponse, completed cases or SBE cases.
TOTAL INCOMPLETE	Total number of remaining incomplete cases in the RCC.



2020 Census: Group Quarters Enumeration Progress & Cost

Group Quarters Enumeration Progress*					
Initial Workload	GQs Added	Total Workload	Completed & Closed Cases	Current Workload	Percent Completed & Closed
195,656	13,345	209,001	137,361	71,640	65.72%

*Only includes the GQE eResponse and GQE in-field sub-operations.

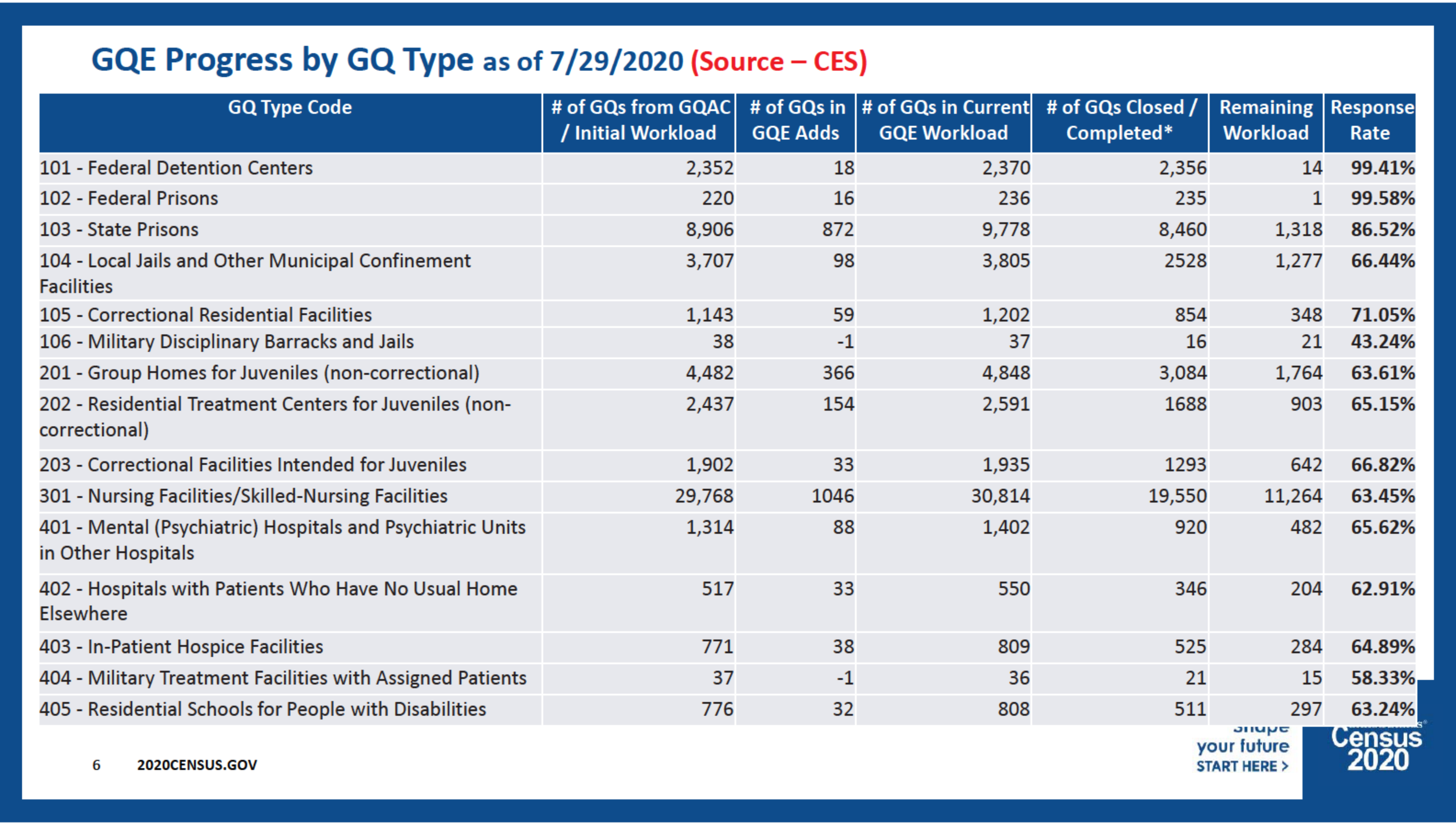
Maritime Vessels Enumeration Progress					
Initial Workload	GQs Added	Total Workload	Vessel Location Reports Checked Into ATAC	Current Workload	Percent Vessel Location Reports Checked Into ATAC
1,420	11	1,431	852	579	59.54%

Costs for Group Quarters Enumeration

Total GQE Budget: $72.4M
Actual GQE Cost: $33.9M

Source: Census Data Lake

DRB Approval Number: CBDRB-FY21-DSEP-002



GQE Progress by GQ Type as of 7/29/2020 (Source – CES)

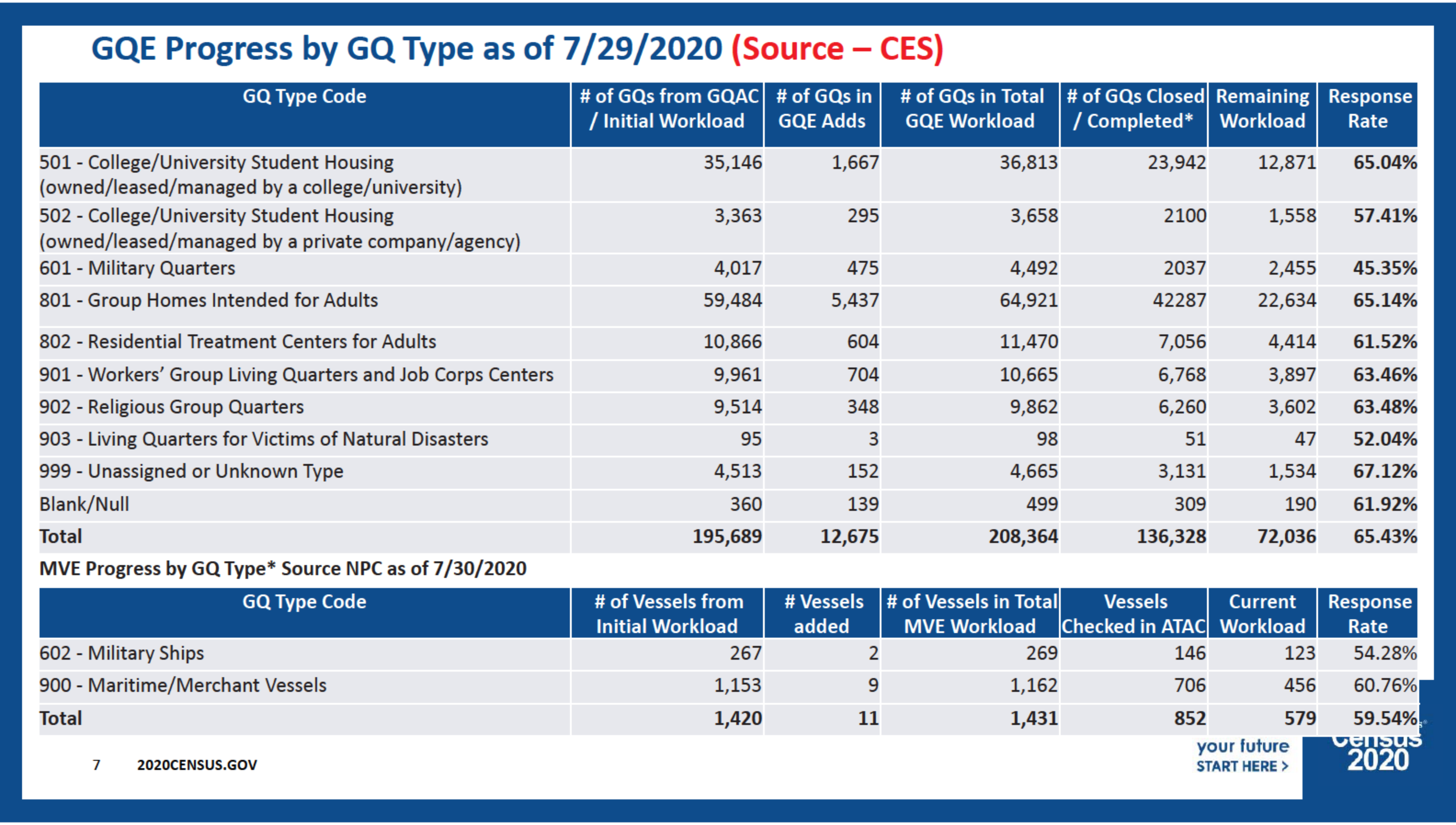
GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Current GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
101 - Federal Detention Centers	2,352	18	2,370	2,356	14	**99.41%**
102 - Federal Prisons	220	16	236	235	1	**99.58%**
103 - State Prisons	8,906	872	9,778	8,460	1,318	**86.52%**
104 - Local Jails and Other Municipal Confinement Facilities	3,707	98	3,805	2528	1,277	**66.44%**
105 - Correctional Residential Facilities	1,143	59	1,202	854	348	**71.05%**
106 - Military Disciplinary Barracks and Jails	38	-1	37	16	21	**43.24%**
201 - Group Homes for Juveniles (non-correctional)	4,482	366	4,848	3,084	1,764	**63.61%**
202 - Residential Treatment Centers for Juveniles (non-correctional)	2,437	154	2,591	1688	903	**65.15%**
203 - Correctional Facilities Intended for Juveniles	1,902	33	1,935	1293	642	**66.82%**
301 - Nursing Facilities/Skilled-Nursing Facilities	29,768	1046	30,814	19,550	11,264	**63.45%**
401 - Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals	1,314	88	1,402	920	482	**65.62%**
402 - Hospitals with Patients Who Have No Usual Home Elsewhere	517	33	550	346	204	**62.91%**
403 - In-Patient Hospice Facilities	771	38	809	525	284	**64.89%**
404 - Military Treatment Facilities with Assigned Patients	37	-1	36	21	15	**58.33%**
405 - Residential Schools for People with Disabilities	776	32	808	511	297	**63.24%**

6 **2020CENSUS.GOV**

Shape your future
START HERE >

Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



GQE Progress by GQ Type as of 7/29/2020 (Source – CES)

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Total GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
501 - College/University Student Housing (owned/leased/managed by a college/university)	35,146	1,667	36,813	23,942	12,871	**65.04%**
502 - College/University Student Housing (owned/leased/managed by a private company/agency)	3,363	295	3,658	2100	1,558	**57.41%**
601 - Military Quarters	4,017	475	4,492	2037	2,455	**45.35%**
801 - Group Homes Intended for Adults	59,484	5,437	64,921	42287	22,634	**65.14%**
802 - Residential Treatment Centers for Adults	10,866	604	11,470	7,056	4,414	**61.52%**
901 - Workers' Group Living Quarters and Job Corps Centers	9,961	704	10,665	6,768	3,897	**63.46%**
902 - Religious Group Quarters	9,514	348	9,862	6,260	3,602	**63.48%**
903 - Living Quarters for Victims of Natural Disasters	95	3	98	51	47	**52.04%**
999 - Unassigned or Unknown Type	4,513	152	4,665	3,131	1,534	**67.12%**
Blank/Null	360	139	499	309	190	**61.92%**
Total	**195,689**	**12,675**	**208,364**	**136,328**	**72,036**	**65.43%**

MVE Progress by GQ Type* Source NPC as of 7/30/2020

GQ Type Code	# of Vessels from Initial Workload	# Vessels added	# of Vessels in Total MVE Workload	Vessels Checked in ATAC	Current Workload	Response Rate
602 - Military Ships	267	2	269	146	123	54.28%
900 - Maritime/Merchant Vessels	1,153	9	1,162	706	456	60.76%
Total	**1,420**	**11**	**1,431**	**852**	**579**	**59.54%**

Issues/ Exceptions/Extensions

Based on guidelines given on June 24 for Restarting operations: "-- *GQE field work starts on July 1, as you well know. The decision is to proceed with GQs that we speak with via phone and who are willing to have us come out, beginning July 1. For GQs that we are not able to contact by phone, we will not proceed to the field until we start NRFU in those states (those states will be updated every Thursday at CIG, as I mentioned above). For these cases with no phone contact, the earliest field visit would be July 16."*

- **As of 7/30, 47 ACOs can make personal visit follow-up on Unresolved cases or cases without appointments. All other ACOs can only visit GQs that have scheduled appointments.**
- **Adjusted finish date for DVS data collection due to several concerns:**
  - Date/timing needed by The Salvation Army (SA) to put out their Directives. While DVS started July 6, SA requested no contact prior to July 22 to allow Directives to be delivered to all of their entities to assure cooperation and knowledge of preferred enumeration method (drop off/pick up).
  - Bureau of Justice requested additional documentation be sent to State Coalitions sharing participation details for the 2020 Census. They requested an extension of the deadline. Deadline extended to August 26.

Backup

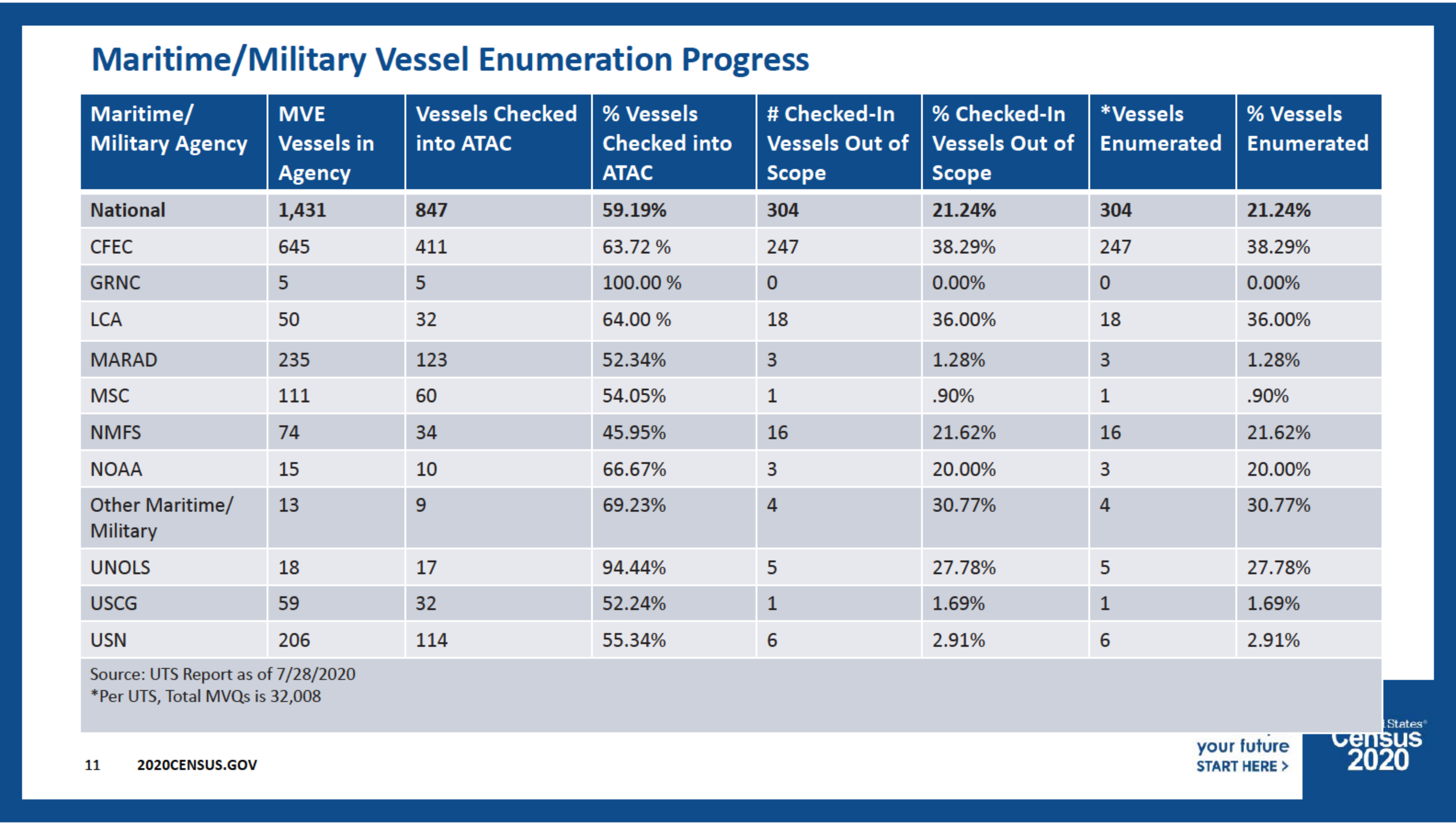
Schedules and Systems

DRB Approval Number: CBDRB-FY21-DSEP-002

GQE Progress: Centurion Referrals /Pending Submissions/Paper Listings

Centurion Referral Paper Listing (non-standard formats)	Pending Submissions in Centurion	Multiple GQs Using One Census ID
As of July 27, NPC received total of 823 Centurion referrals • 508 had completed keying • 83 issues pending, issues escalated to DCMD. • 50 where some coming in from GQ Admins or individual type submittals with one ICQ. • 26 are in progress, being currently worked.	CR submitted to have the remaining pending cases pushed out on 8/3. • 384 as of 7/28/2020	As FLD and DCMD hears from GQ admins that cases were already submitted, an outreach goes to DSSD for confirmation. This process has been successful in identifying GQs that fit in this category.

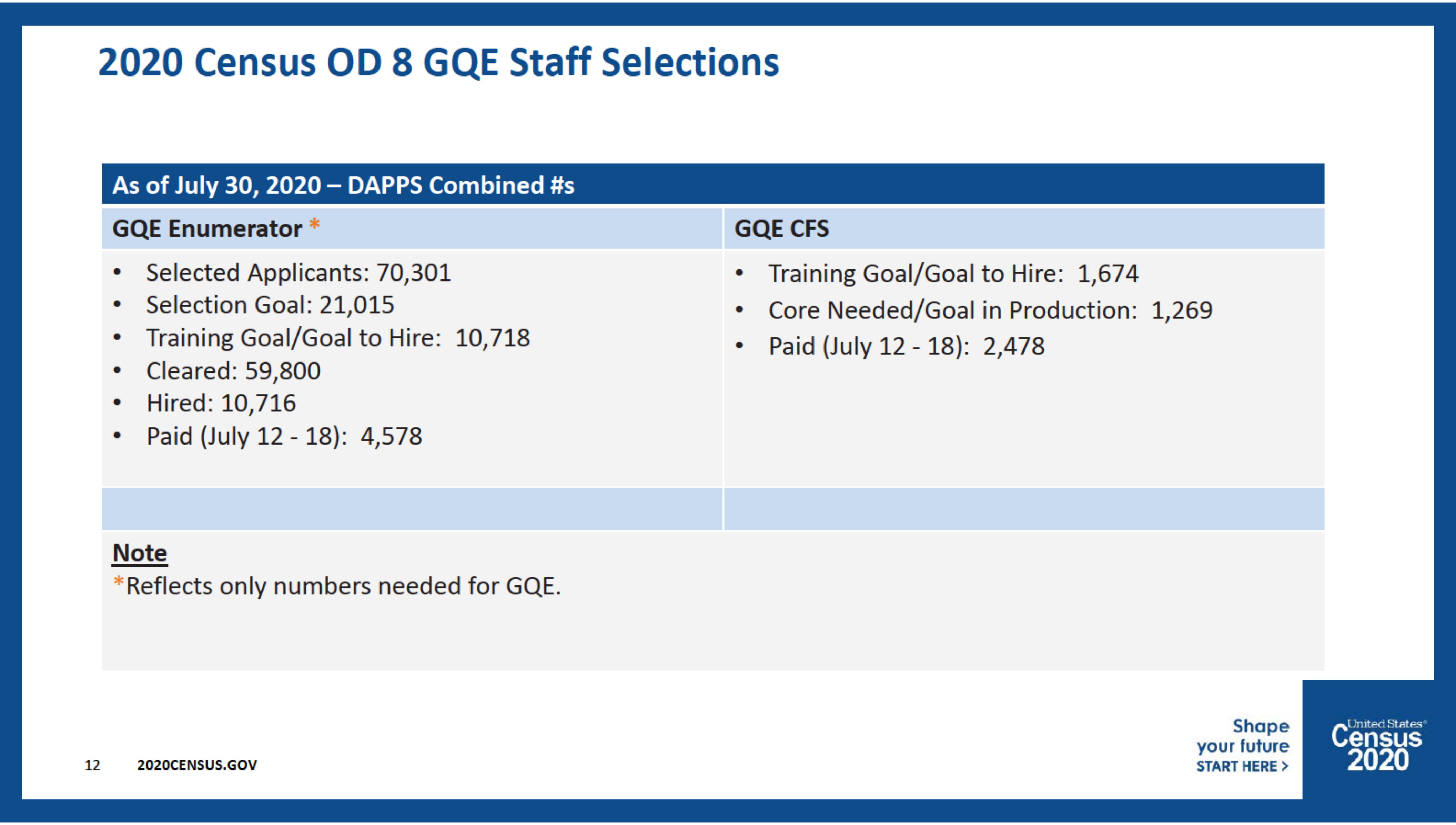
DRB Approval Number: CBDRB-FY21-DSEP-002



Maritime/Military Vessel Enumeration Progress

Maritime/ Military Agency	MVE Vessels in Agency	Vessels Checked into ATAC	% Vessels Checked into ATAC	# Checked-In Vessels Out of Scope	% Checked-In Vessels Out of Scope	*Vessels Enumerated	% Vessels Enumerated
National	**1,431**	**847**	**59.19%**	**304**	**21.24%**	**304**	**21.24%**
CFEC	645	411	63.72 %	247	38.29%	247	38.29%
GRNC	5	5	100.00 %	0	0.00%	0	0.00%
LCA	50	32	64.00 %	18	36.00%	18	36.00%
MARAD	235	123	52.34%	3	1.28%	3	1.28%
MSC	111	60	54.05%	1	.90%	1	.90%
NMFS	74	34	45.95%	16	21.62%	16	21.62%
NOAA	15	10	66.67%	3	20.00%	3	20.00%
Other Maritime/ Military	13	9	69.23%	4	30.77%	4	30.77%
UNOLS	18	17	94.44%	5	27.78%	5	27.78%
USCG	59	32	52.24%	1	1.69%	1	1.69%
USN	206	114	55.34%	6	2.91%	6	2.91%

Source: UTS Report as of 7/28/2020
*Per UTS, Total MVQs is 32,008



2020 Census OD 8 GQE Staff Selections

As of July 30, 2020 – DAPPS Combined #s

GQE Enumerator *	GQE CFS
• Selected Applicants: 70,301 • Selection Goal: 21,015 • Training Goal/Goal to Hire: 10,718 • Cleared: 59,800 • Hired: 10,716 • Paid (July 12 - 18): 4,578	• Training Goal/Goal to Hire: 1,674 • Core Needed/Goal in Production: 1,269 • Paid (July 12 - 18): 2,478

Note

*Reflects only numbers needed for GQE.

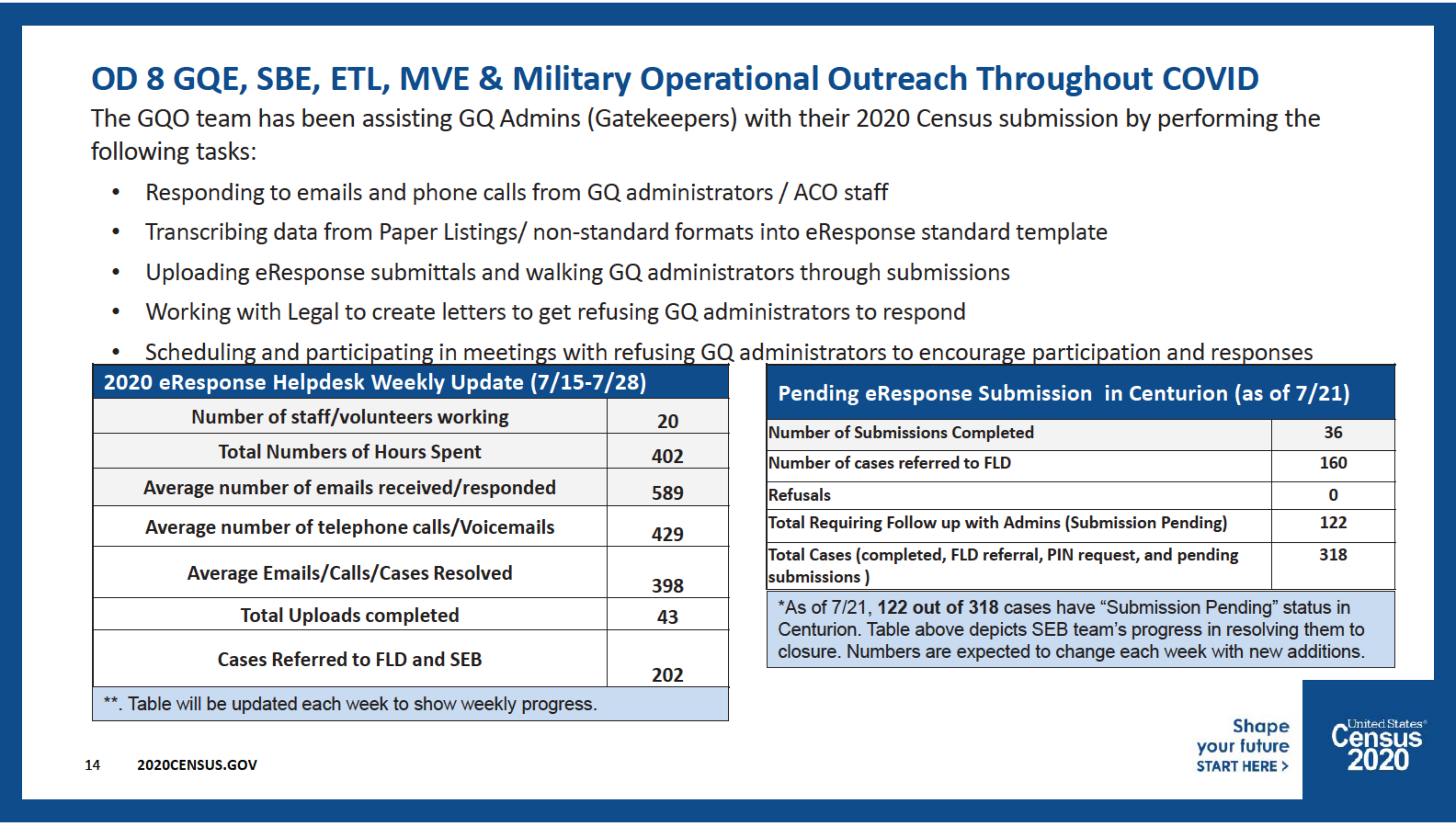
Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

OD 8 GQE Milestone Conduct & Training Dates

Conduct Activity	Proposed Start	Proposed Finish
Conduct MVE Operation	04/01/20 (A)	07/31/20 (P)
Conduct GQE eResponse Operation	04/01/20 (A)	08/07/20 (P)
Conduct GQE Operation Field	04/20/20 (A)	08/26/20 (P)
Conduct GQE In-Person Operation	07/01/20 (A)	08/26/20 (P)
Conduct GQ Reinterview Operation	07/02/20 (A)	09/03/20 (P)
Conduct DVS Enumeration Field	07/06/20 (A)	08/26/20 (P)



OD 8 GQE, SBE, ETL, MVE & Military Operational Outreach Throughout COVID

The GQO team has been assisting GQ Admins (Gatekeepers) with their 2020 Census submission by performing the following tasks:

- Responding to emails and phone calls from GQ administrators / ACO staff
- Transcribing data from Paper Listings/ non-standard formats into eResponse standard template
- Uploading eResponse submittals and walking GQ administrators through submissions
- Working with Legal to create letters to get refusing GQ administrators to respond
- Scheduling and participating in meetings with refusing GQ administrators to encourage participation and responses

2020 eResponse Helpdesk Weekly Update (7/15-7/28)	
Number of staff/volunteers working	20
Total Numbers of Hours Spent	402
Average number of emails received/responded	589
Average number of telephone calls/Voicemails	429
Average Emails/Calls/Cases Resolved	398
Total Uploads completed	43
Cases Referred to FLD and SEB	202

**. Table will be updated each week to show weekly progress.

Pending eResponse Submission in Centurion (as of 7/21)	
Number of Submissions Completed	36
Number of cases referred to FLD	160
Refusals	0
Total Requiring Follow up with Admins (Submission Pending)	122
Total Cases (completed, FLD referral, PIN request, and pending submissions)	318

*As of 7/21, **122 out of 318** cases have "Submission Pending" status in Centurion. Table above depicts SEB team's progress in resolving them to closure. Numbers are expected to change each week with new additions.

Shape
your future
START HERE >

14 2020CENSUS.GOV

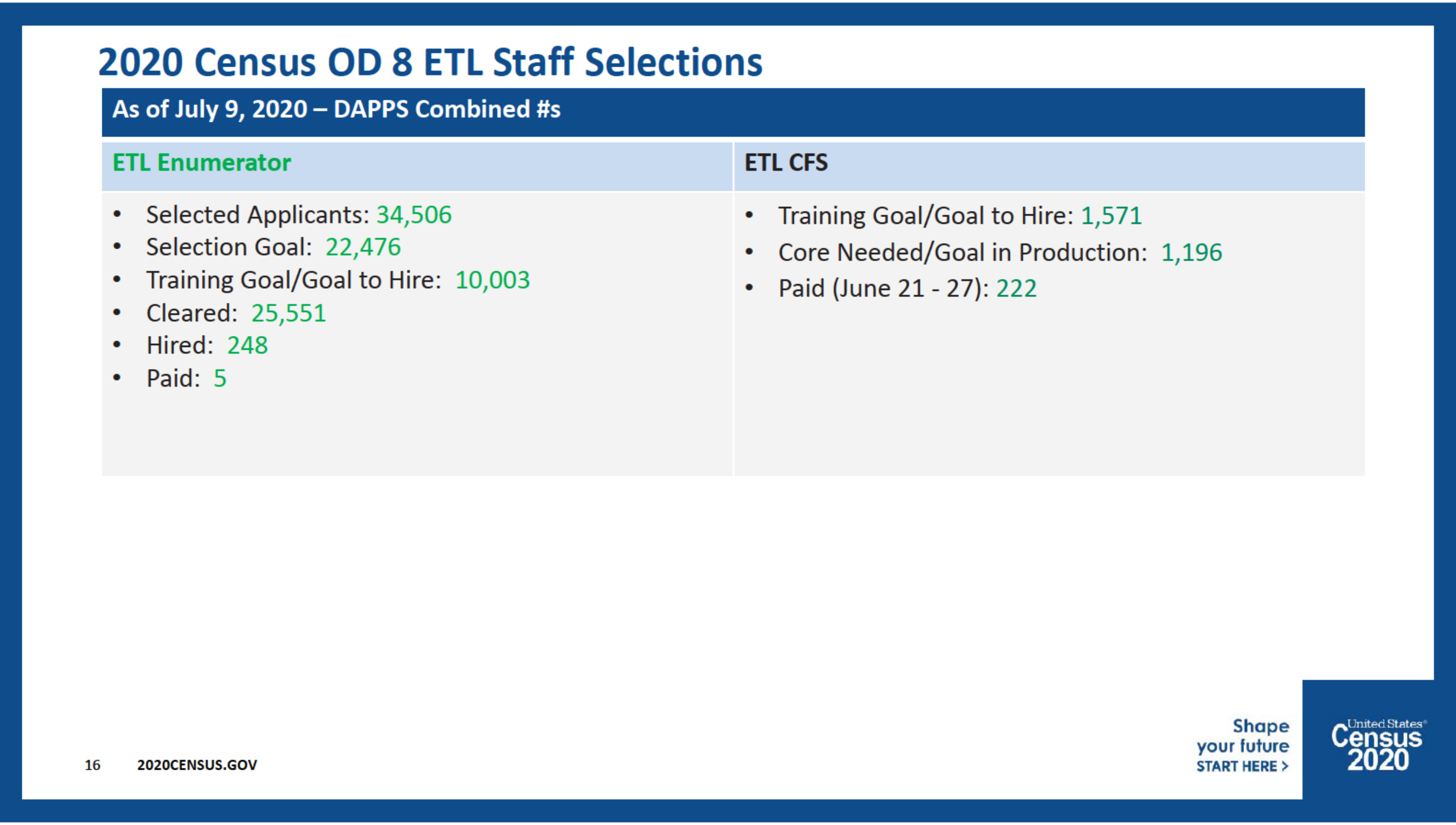
OD 8 GQE, SBE, ETL, MVE & Military System Updates

- **Box Check In Issue Resolution**
  - Box Check In working group stood up with members from NPC, FOCS, CDL, UTS, DCMD, FLD, TI and DSSD to address box check in issues (which prevent cases from reaching completion status) and drive them to closure
- **Field OCS Updates**
  - CR 1853 approved which allows for more flexibility to allow cases that are currently stuck in a certain status due to user error to move forward
- **FACO**
  - As of 7/30 per UTS 100/108 received = 93 % complete. On going meetings with agencies to discuss data anomalies.
- **Military**
  - MOB continues to work with the military reps from the CJSWG to get POC updates and resolve issues with base access.

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census OD 8 ETL Staff Selections

As of July 9, 2020 – DAPPS Combined #s

ETL Enumerator

- Selected Applicants: 34,506
- Selection Goal: 22,476
- Training Goal/Goal to Hire: 10,003
- Cleared: 25,551
- Hired: 248
- Paid: 5

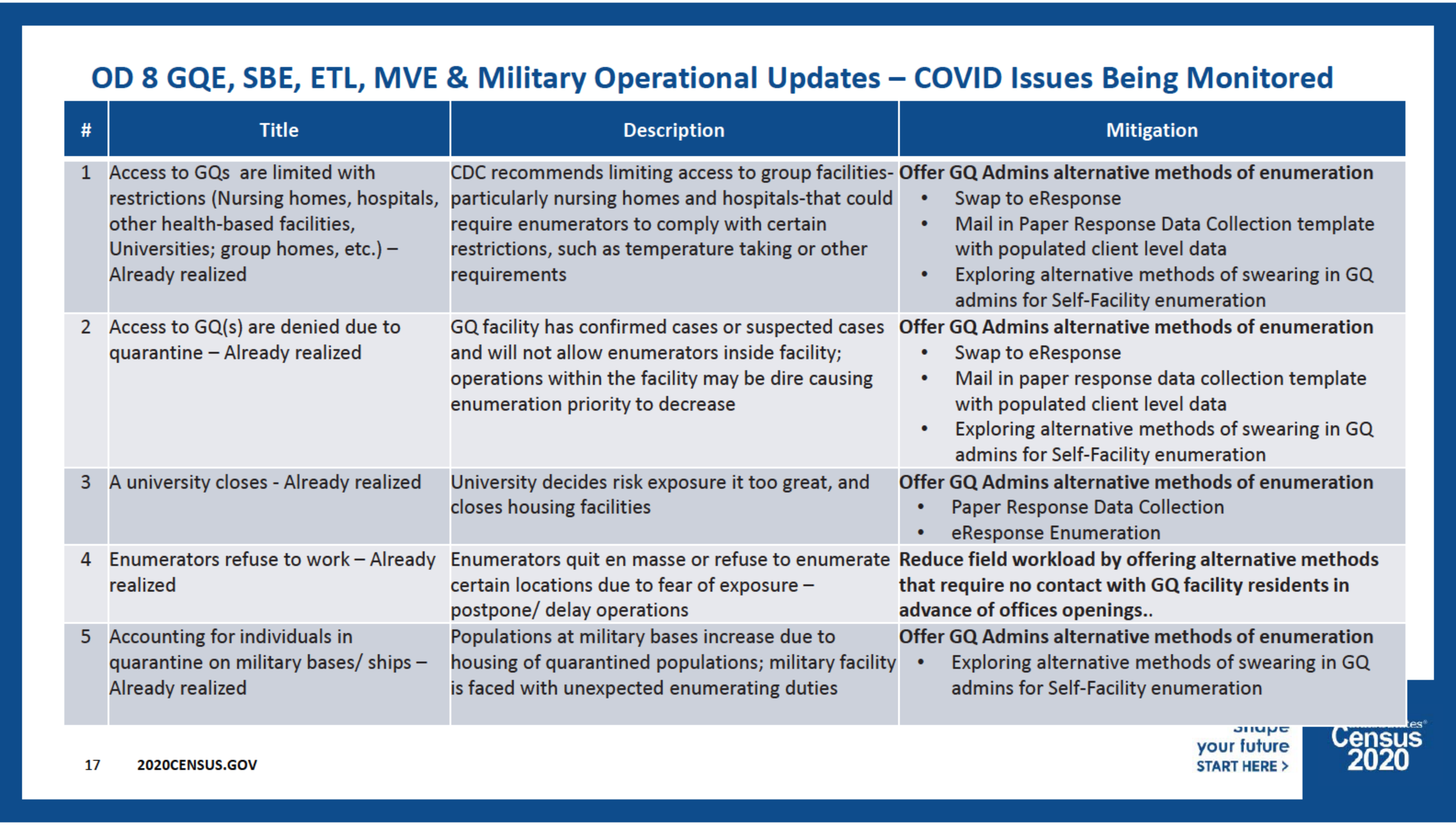
ETL CFS

- Training Goal/Goal to Hire: 1,571
- Core Needed/Goal in Production: 1,196
- Paid (June 21 - 27): 222

Shape your future START HERE >

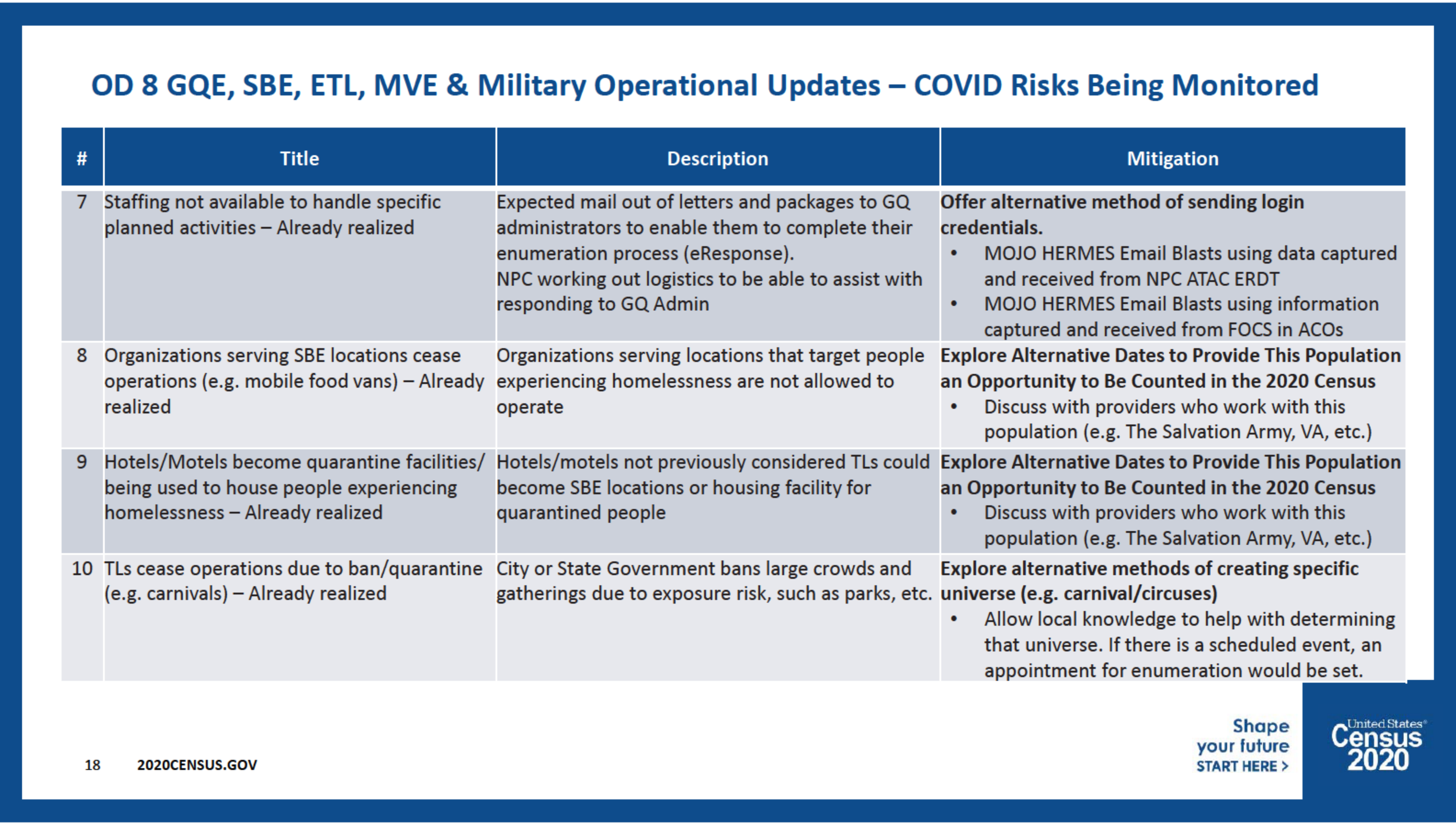
United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Issues Being Monitored

#	Title	Description	Mitigation
1	Access to GQs are limited with restrictions (Nursing homes, hospitals, other health-based facilities, Universities; group homes, etc.) – Already realized	CDC recommends limiting access to group facilities-particularly nursing homes and hospitals-that could require enumerators to comply with certain restrictions, such as temperature taking or other requirements	**Offer GQ Admins alternative methods of enumeration** • Swap to eResponse • Mail in Paper Response Data Collection template with populated client level data • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration
2	Access to GQ(s) are denied due to quarantine – Already realized	GQ facility has confirmed cases or suspected cases and will not allow enumerators inside facility; operations within the facility may be dire causing enumeration priority to decrease	**Offer GQ Admins alternative methods of enumeration** • Swap to eResponse • Mail in paper response data collection template with populated client level data • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration
3	A university closes - Already realized	University decides risk exposure it too great, and closes housing facilities	**Offer GQ Admins alternative methods of enumeration** • Paper Response Data Collection • eResponse Enumeration
4	Enumerators refuse to work – Already realized	Enumerators quit en masse or refuse to enumerate certain locations due to fear of exposure – postpone/ delay operations	**Reduce field workload by offering alternative methods that require no contact with GQ facility residents in advance of offices openings..**
5	Accounting for individuals in quarantine on military bases/ ships – Already realized	Populations at military bases increase due to housing of quarantined populations; military facility is faced with unexpected enumerating duties	**Offer GQ Admins alternative methods of enumeration** • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration



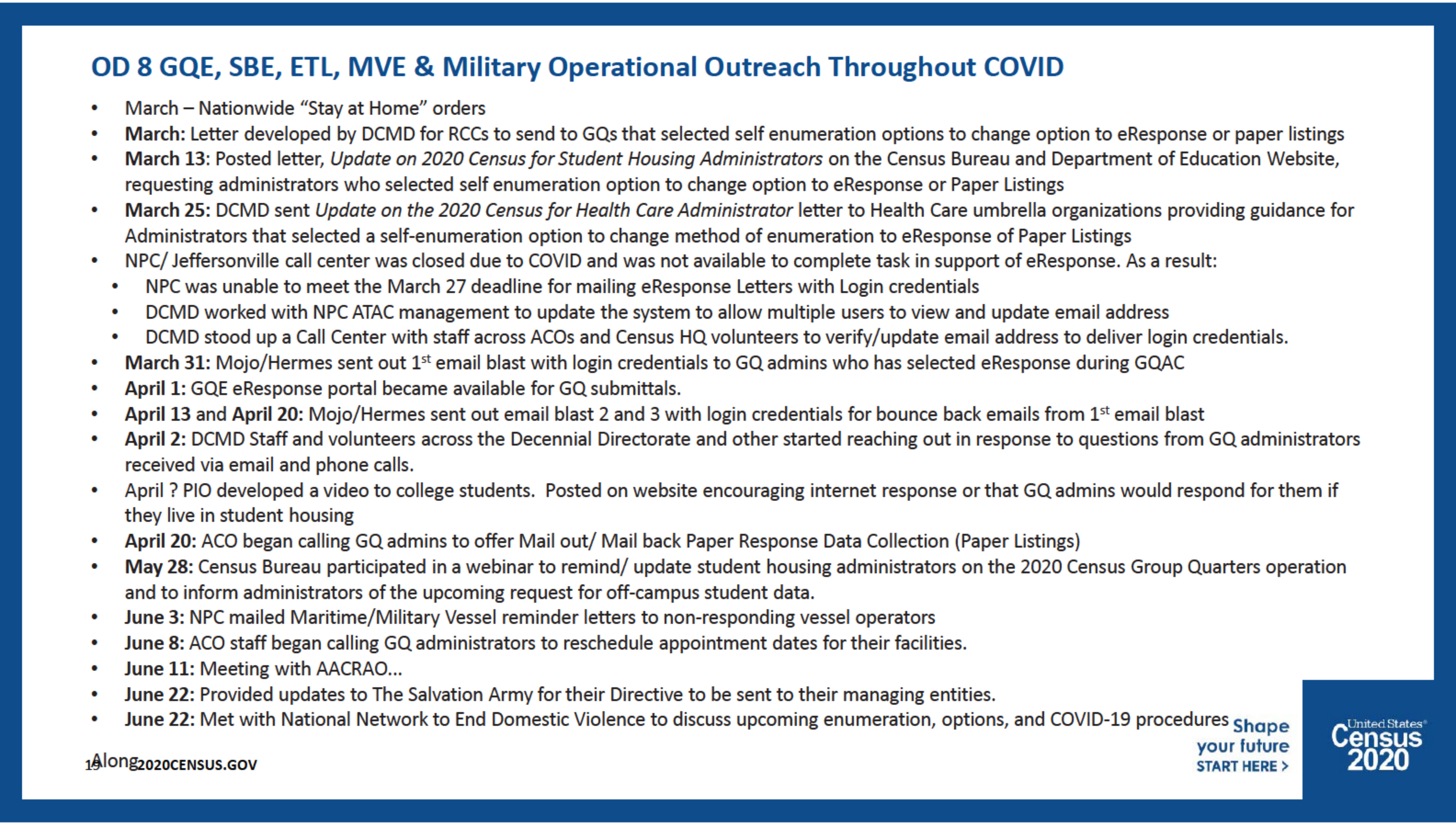
OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Risks Being Monitored

#	Title	Description	Mitigation
7	Staffing not available to handle specific planned activities – Already realized	Expected mail out of letters and packages to GQ administrators to enable them to complete their enumeration process (eResponse). NPC working out logistics to be able to assist with responding to GQ Admin	**Offer alternative method of sending login credentials.** • MOJO HERMES Email Blasts using data captured and received from NPC ATAC ERDT • MOJO HERMES Email Blasts using information captured and received from FOCS in ACOs
8	Organizations serving SBE locations cease operations (e.g. mobile food vans) – Already realized	Organizations serving locations that target people experiencing homelessness are not allowed to operate	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
9	Hotels/Motels become quarantine facilities/ being used to house people experiencing homelessness – Already realized	Hotels/motels not previously considered TLs could become SBE locations or housing facility for quarantined people	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
10	TLs cease operations due to ban/quarantine (e.g. carnivals) – Already realized	City or State Government bans large crowds and gatherings due to exposure risk, such as parks, etc.	**Explore alternative methods of creating specific universe (e.g. carnival/circuses)** • Allow local knowledge to help with determining that universe. If there is a scheduled event, an appointment for enumeration would be set.

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

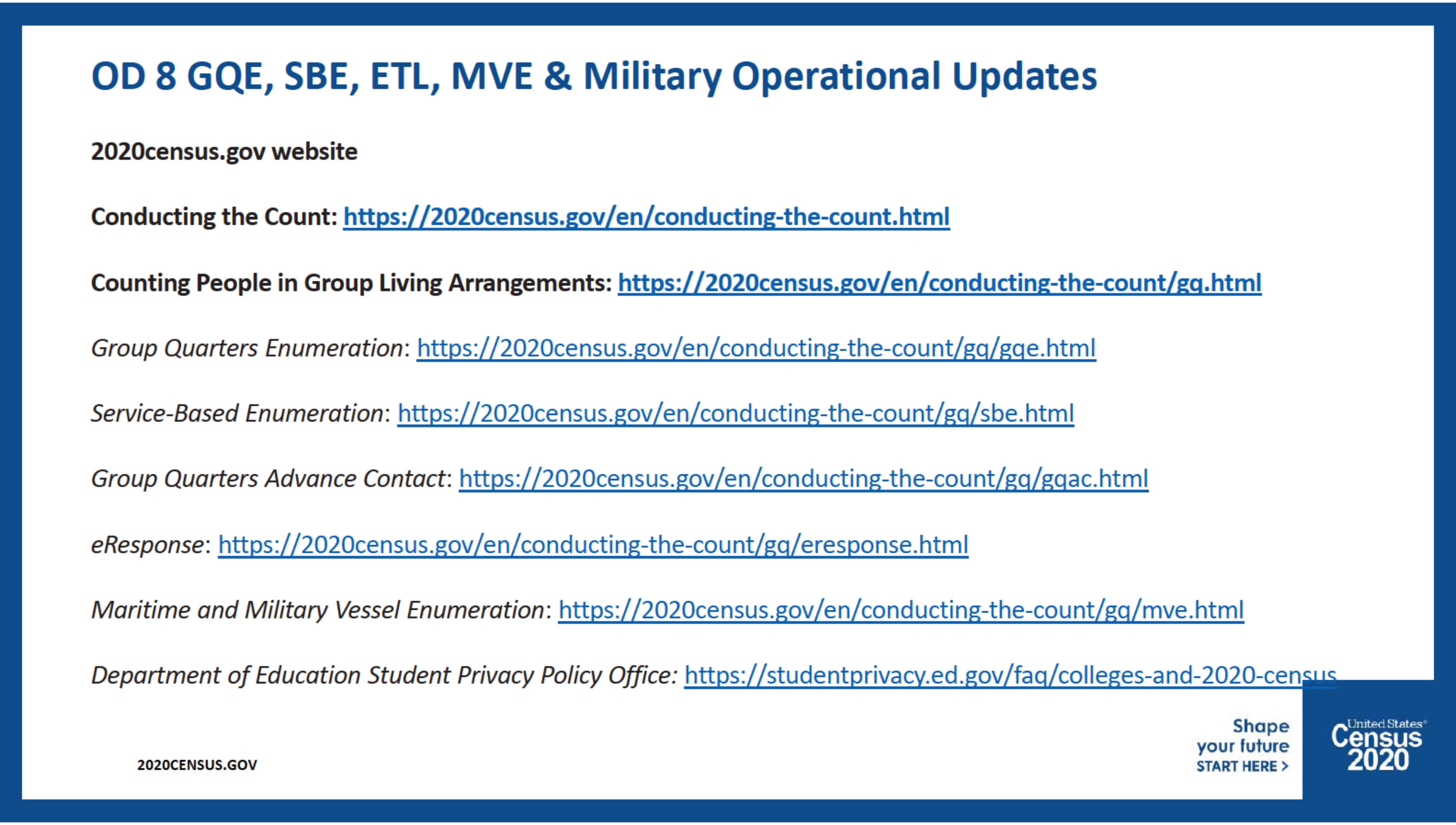


OD 8 GQE, SBE, ETL, MVE & Military Operational Outreach Throughout COVID

- March – Nationwide “Stay at Home” orders
- **March:** Letter developed by DCMD for RCCs to send to GQs that selected self enumeration options to change option to eResponse or paper listings
- **March 13:** Posted letter, *Update on 2020 Census for Student Housing Administrators* on the Census Bureau and Department of Education Website, requesting administrators who selected self enumeration option to change option to eResponse or Paper Listings
- **March 25:** DCMD sent *Update on the 2020 Census for Health Care Administrator* letter to Health Care umbrella organizations providing guidance for Administrators that selected a self-enumeration option to change method of enumeration to eResponse of Paper Listings
- NPC/ Jeffersonville call center was closed due to COVID and was not available to complete task in support of eResponse. As a result:
  - NPC was unable to meet the March 27 deadline for mailing eResponse Letters with Login credentials
  - DCMD worked with NPC ATAC management to update the system to allow multiple users to view and update email address
  - DCMD stood up a Call Center with staff across ACOs and Census HQ volunteers to verify/update email address to deliver login credentials.
- **March 31:** Mojo/Hermes sent out 1st email blast with login credentials to GQ admins who has selected eResponse during GQAC
- **April 1:** GQE eResponse portal became available for GQ submittals.
- **April 13** and **April 20:** Mojo/Hermes sent out email blast 2 and 3 with login credentials for bounce back emails from 1st email blast
- **April 2:** DCMD Staff and volunteers across the Decennial Directorate and other started reaching out in response to questions from GQ administrators received via email and phone calls.
- April ? PIO developed a video to college students. Posted on website encouraging internet response or that GQ admins would respond for them if they live in student housing
- **April 20:** ACO began calling GQ admins to offer Mail out/ Mail back Paper Response Data Collection (Paper Listings)
- **May 28:** Census Bureau participated in a webinar to remind/ update student housing administrators on the 2020 Census Group Quarters operation and to inform administrators of the upcoming request for off-campus student data.
- **June 3:** NPC mailed Maritime/Military Vessel reminder letters to non-responding vessel operators
- **June 8:** ACO staff began calling GQ administrators to reschedule appointment dates for their facilities.
- **June 11:** Meeting with AACRAO…
- **June 22:** Provided updates to The Salvation Army for their Directive to be sent to their managing entities.
- **June 22:** Met with National Network to End Domestic Violence to discuss upcoming enumeration, options, and COVID-19 procedures

Shape your future START HERE >

United States Census 2020



OD 8 GQE, SBE, ETL, MVE & Military Operational Updates

2020census.gov website

Conducting the Count: https://2020census.gov/en/conducting-the-count.html

Counting People in Group Living Arrangements: https://2020census.gov/en/conducting-the-count/gq.html

Group Quarters Enumeration: https://2020census.gov/en/conducting-the-count/gq/gqe.html

Service-Based Enumeration: https://2020census.gov/en/conducting-the-count/gq/sbe.html

Group Quarters Advance Contact: https://2020census.gov/en/conducting-the-count/gq/gqac.html

eResponse: https://2020census.gov/en/conducting-the-count/gq/eresponse.html

Maritime and Military Vessel Enumeration: https://2020census.gov/en/conducting-the-count/gq/mve.html

Department of Education Student Privacy Policy Office: https://studentprivacy.ed.gov/faq/colleges-and-2020-census

2020CENSUS.GOV

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

United States Census 2020

More information on the 2020 Census: http://www.census.gov/2020Census

American Community Survey

More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

instagram.com/uscensusbureau

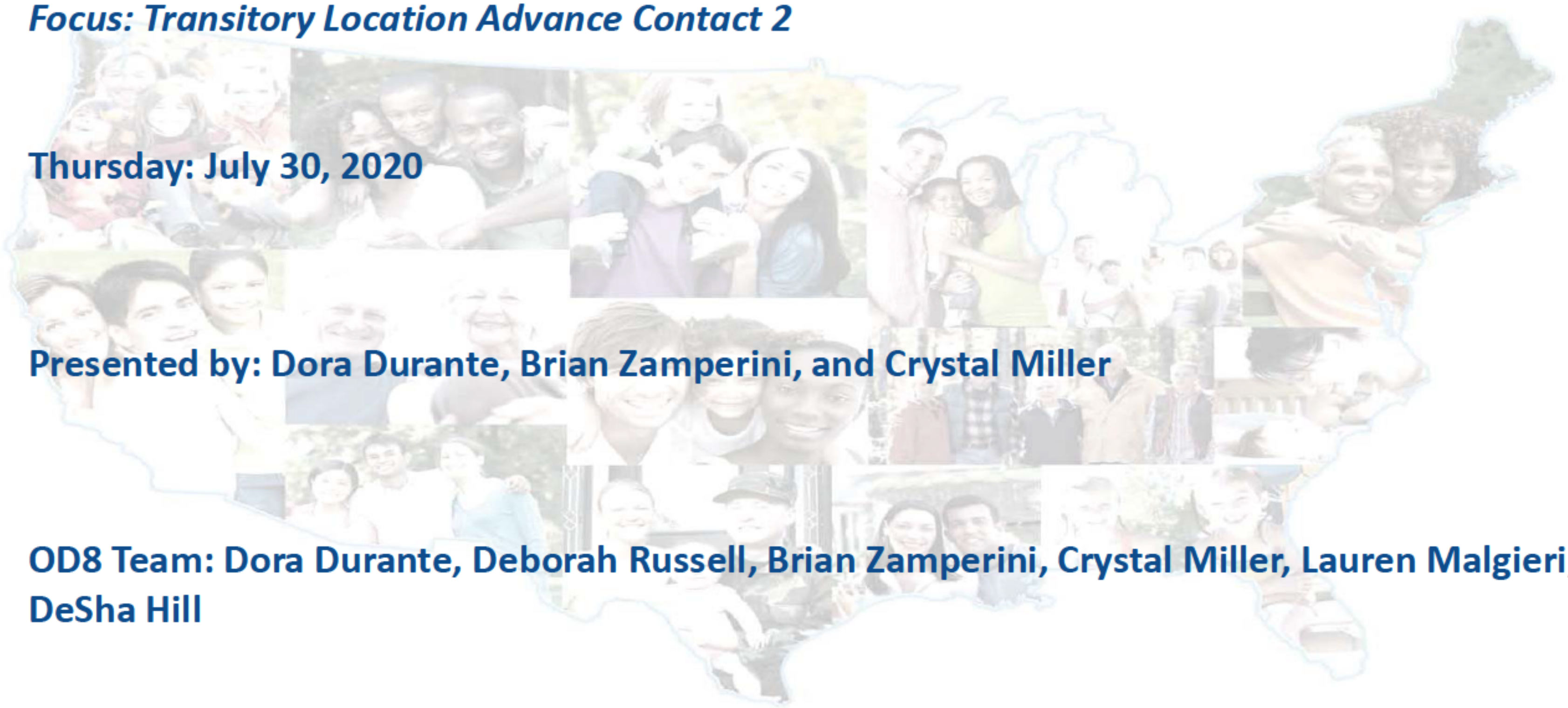
Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Operational Delivery 8: Transitory Locations

Focus: Transitory Location Advance Contact 2

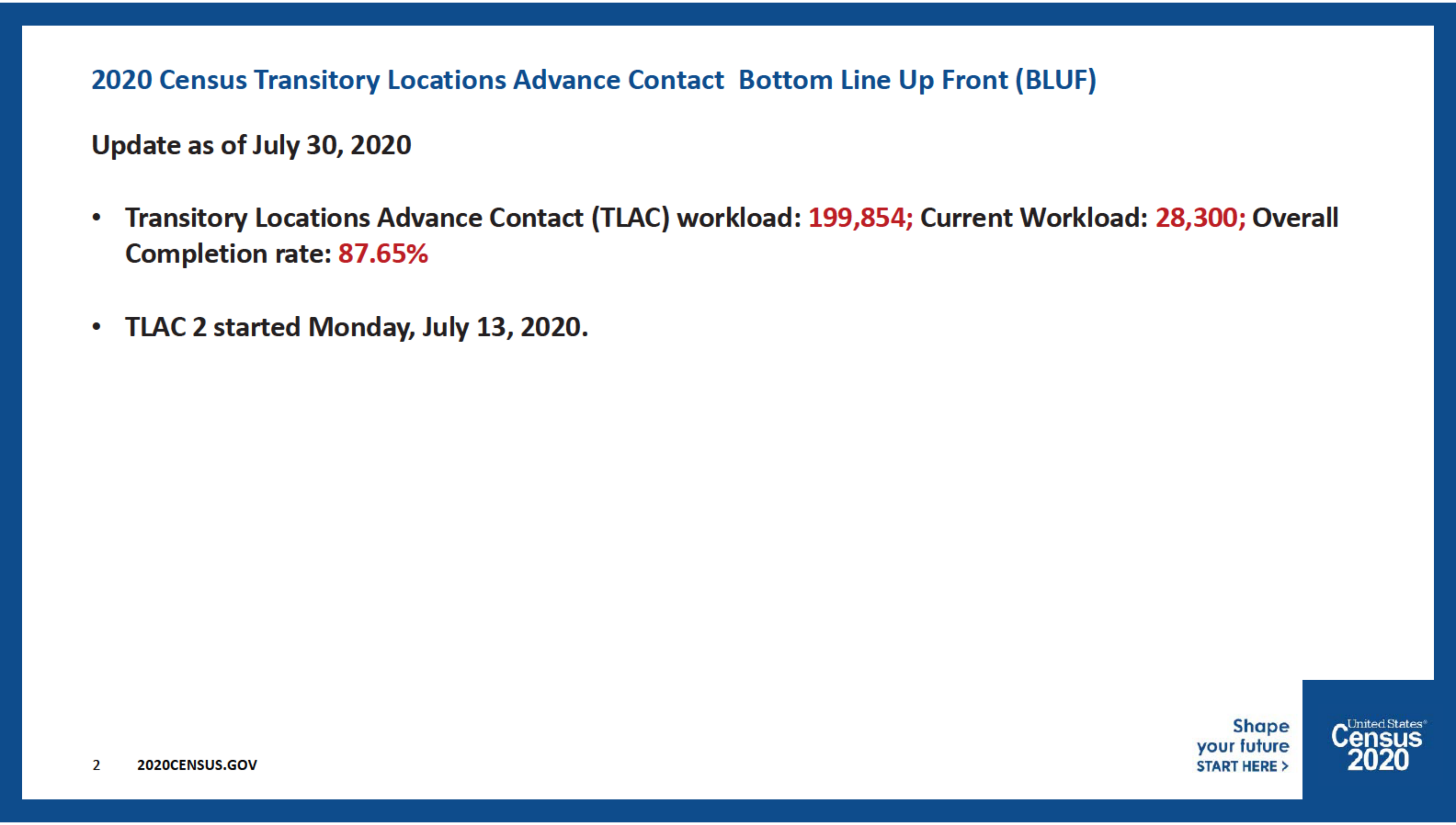


Thursday: July 30, 2020

Presented by: Dora Durante, Brian Zamperini, and Crystal Miller

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

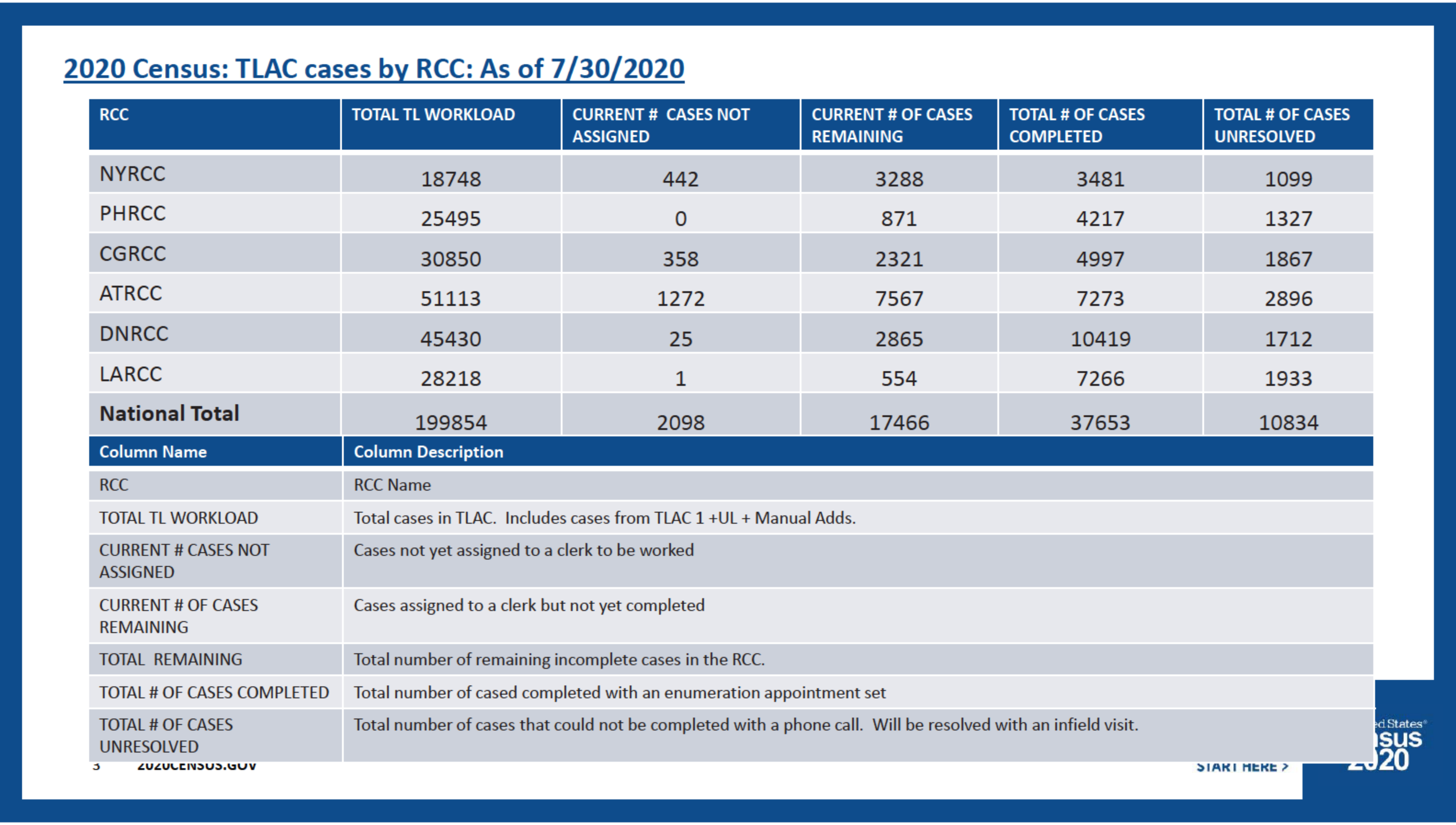
DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census Transitory Locations Advance Contact Bottom Line Up Front (BLUF)

Update as of July 30, 2020

- **Transitory Locations Advance Contact (TLAC) workload: 199,854; Current Workload: 28,300; Overall Completion rate: 87.65%**
- **TLAC 2 started Monday, July 13, 2020.**



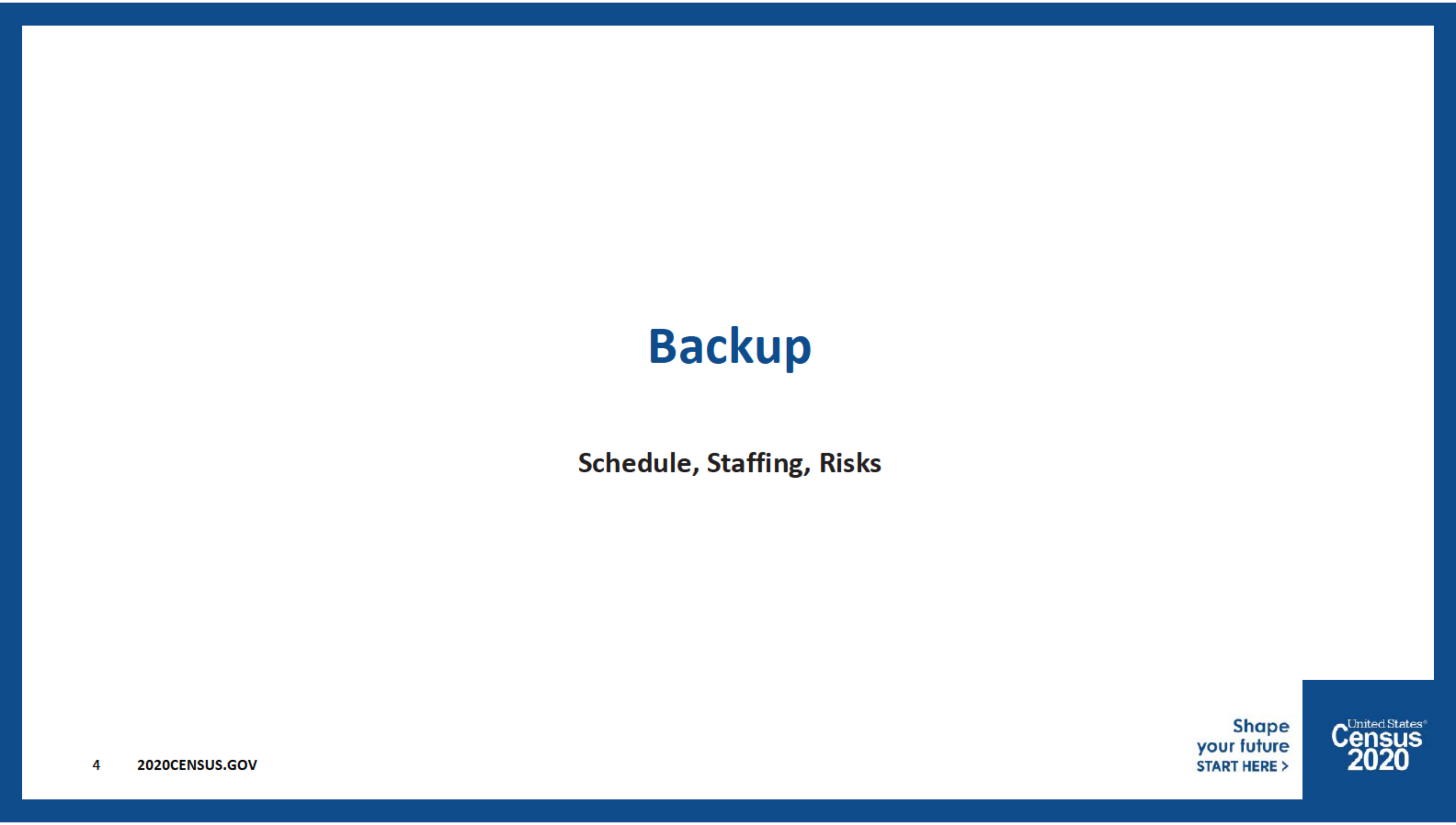
2020 Census: TLAC cases by RCC: As of 7/30/2020

RCC	TOTAL TL WORKLOAD	CURRENT # CASES NOT ASSIGNED	CURRENT # OF CASES REMAINING	TOTAL # OF CASES COMPLETED	TOTAL # OF CASES UNRESOLVED
NYRCC	18748	442	3288	3481	1099
PHRCC	25495	0	871	4217	1327
CGRCC	30850	358	2321	4997	1867
ATRCC	51113	1272	7567	7273	2896
DNRCC	45430	25	2865	10419	1712
LARCC	28218	1	554	7266	1933
National Total	199854	2098	17466	37653	10834

Column Name	Column Description
RCC	RCC Name
TOTAL TL WORKLOAD	Total cases in TLAC. Includes cases from TLAC 1 +UL + Manual Adds.
CURRENT # CASES NOT ASSIGNED	Cases not yet assigned to a clerk to be worked
CURRENT # OF CASES REMAINING	Cases assigned to a clerk but not yet completed
TOTAL REMAINING	Total number of remaining incomplete cases in the RCC.
TOTAL # OF CASES COMPLETED	Total number of cased completed with an enumeration appointment set
TOTAL # OF CASES UNRESOLVED	Total number of cases that could not be completed with a phone call. Will be resolved with an infield visit.

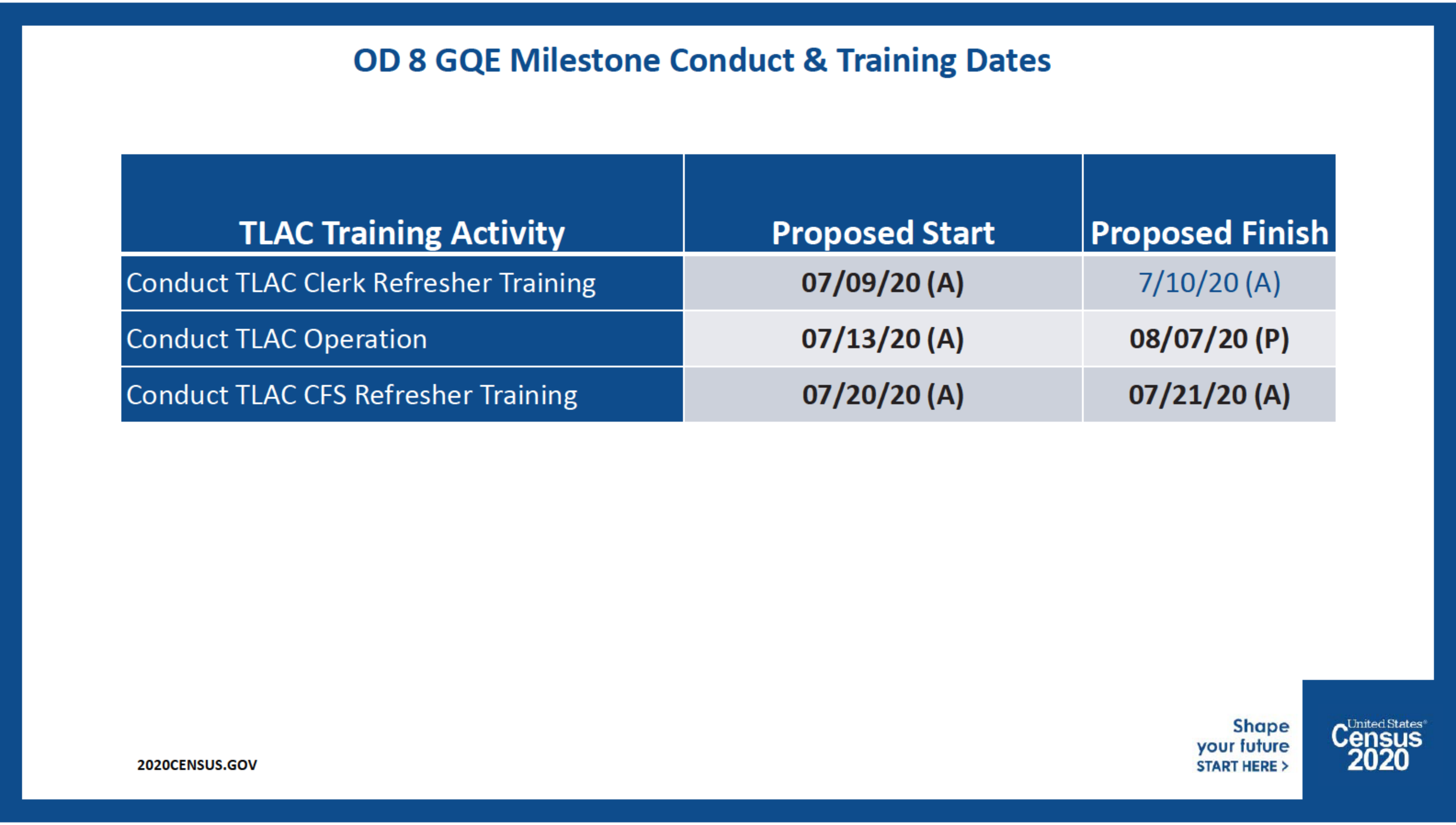
START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



Backup

Schedule, Staffing, Risks

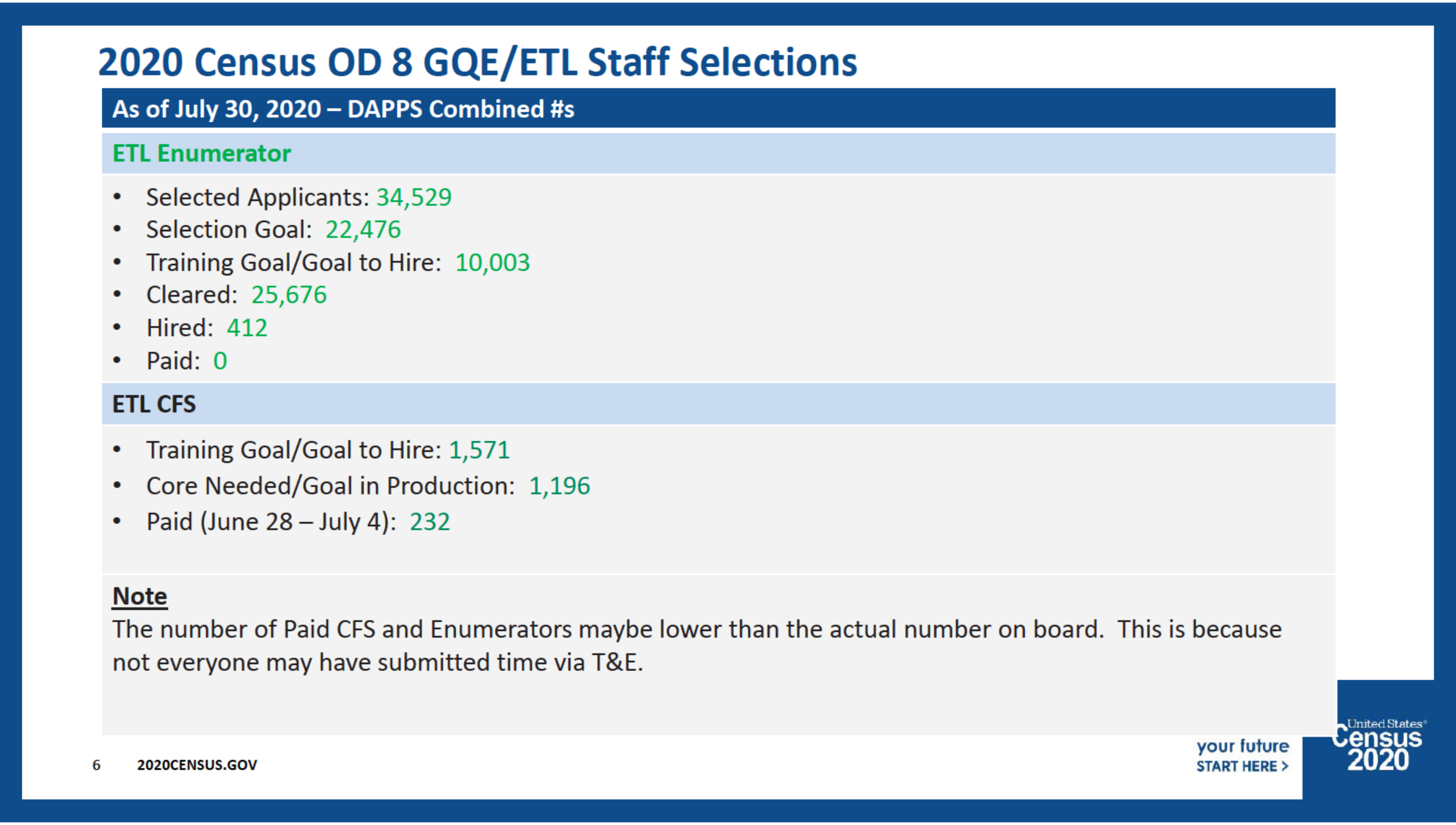


OD 8 GQE Milestone Conduct & Training Dates

TLAC Training Activity	Proposed Start	Proposed Finish
Conduct TLAC Clerk Refresher Training	07/09/20 (A)	7/10/20 (A)
Conduct TLAC Operation	07/13/20 (A)	08/07/20 (P)
Conduct TLAC CFS Refresher Training	07/20/20 (A)	07/21/20 (A)

2020CENSUS.GOV

Shape
your future
START HERE >



2020 Census OD 8 GQE/ETL Staff Selections

As of July 30, 2020 – DAPPS Combined #s

ETL Enumerator

- Selected Applicants: 34,529
- Selection Goal: 22,476
- Training Goal/Goal to Hire: 10,003
- Cleared: 25,676
- Hired: 412
- Paid: 0

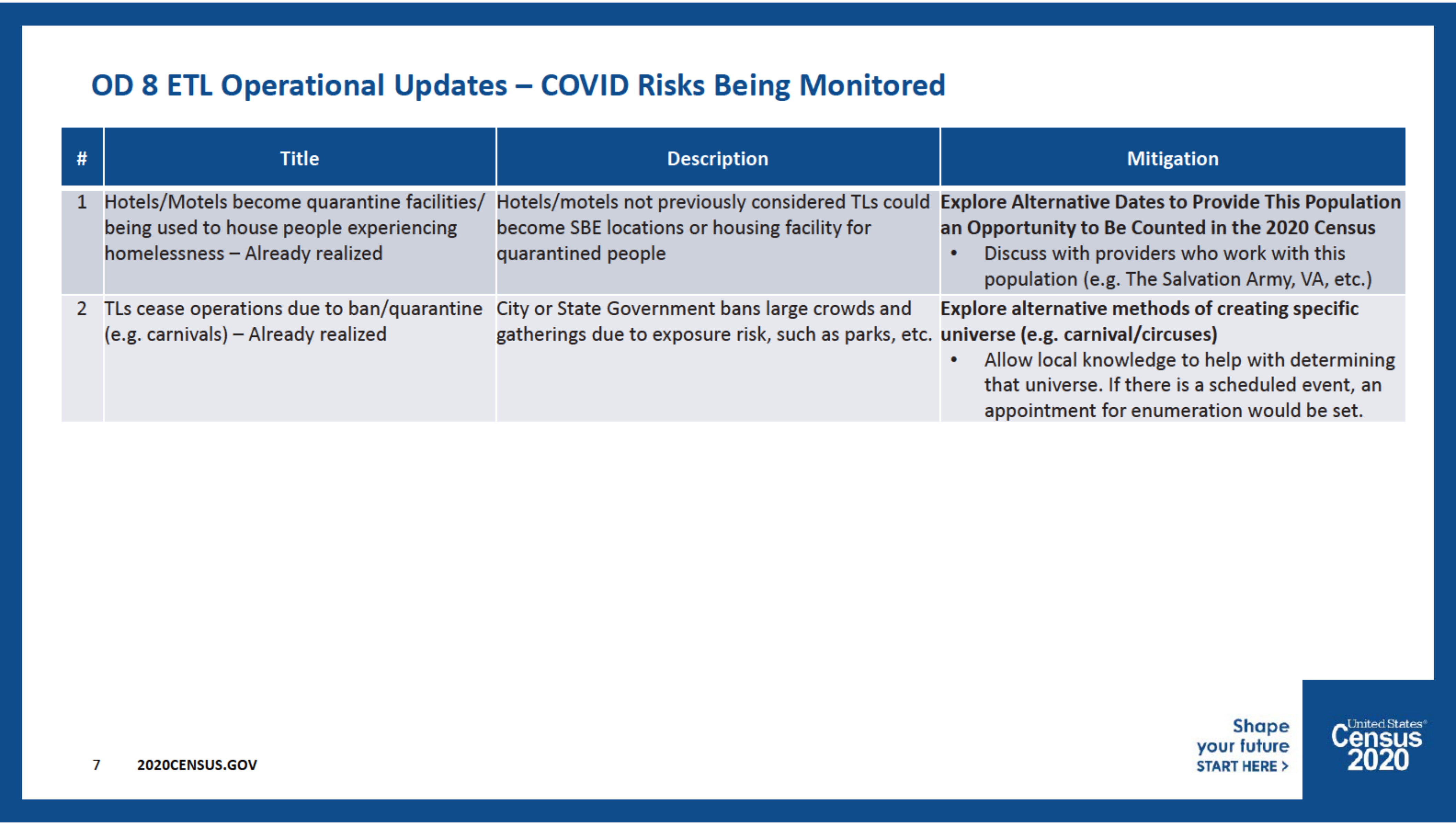
ETL CFS

- Training Goal/Goal to Hire: 1,571
- Core Needed/Goal in Production: 1,196
- Paid (June 28 – July 4): 232

Note

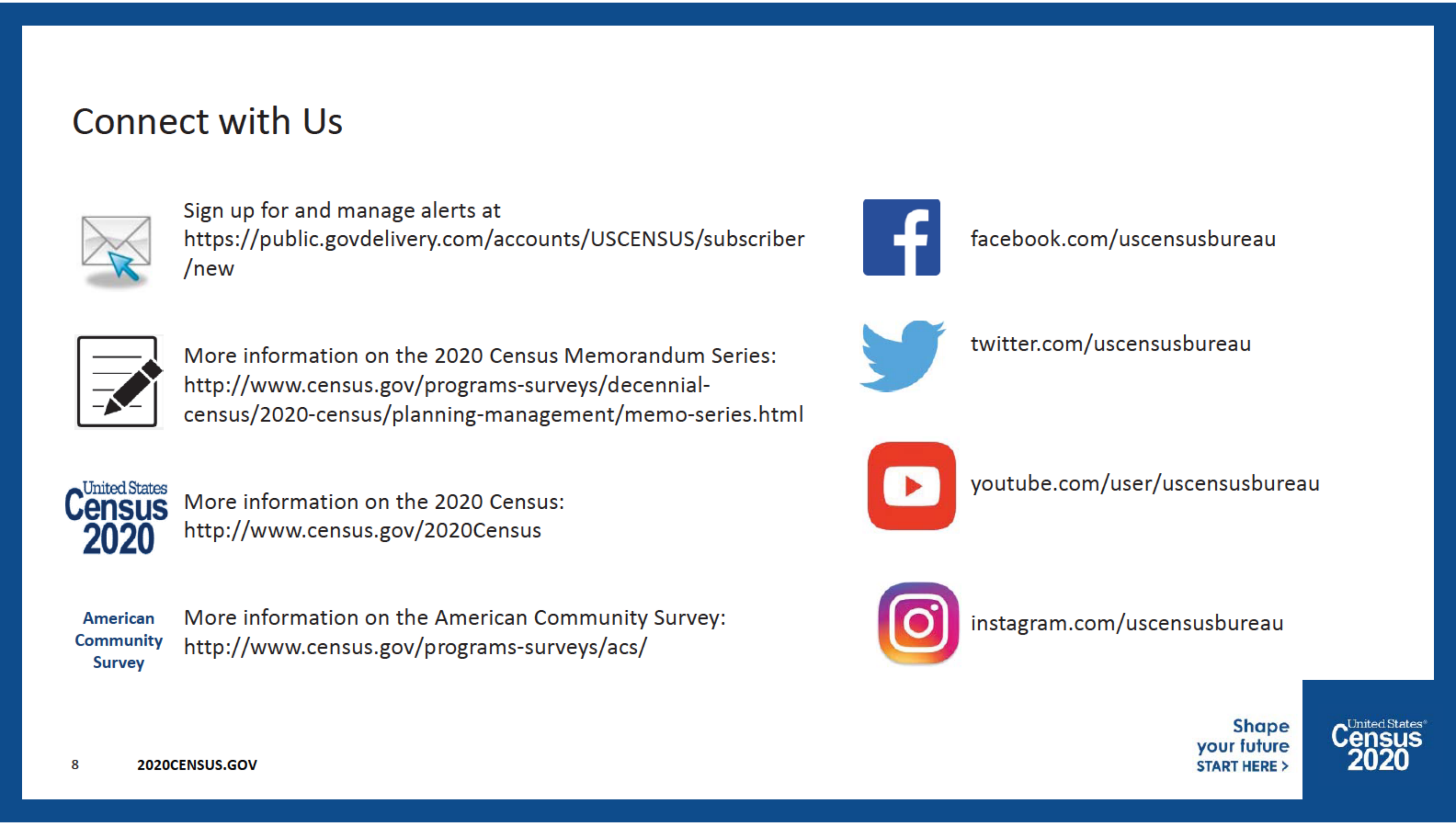
The number of Paid CFS and Enumerators maybe lower than the actual number on board. This is because not everyone may have submitted time via T&E.

DRB Approval Number: CBDRB-FY21-DSEP-002



OD 8 ETL Operational Updates – COVID Risks Being Monitored

#	Title	Description	Mitigation
1	Hotels/Motels become quarantine facilities/ being used to house people experiencing homelessness – Already realized	Hotels/motels not previously considered TLs could become SBE locations or housing facility for quarantined people	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
2	TLs cease operations due to ban/quarantine (e.g. carnivals) – Already realized	City or State Government bans large crowds and gatherings due to exposure risk, such as parks, etc.	**Explore alternative methods of creating specific universe (e.g. carnival/circuses)** • Allow local knowledge to help with determining that universe. If there is a scheduled event, an appointment for enumeration would be set.



Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

More information on the 2020 Census: http://www.census.gov/2020Census

More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

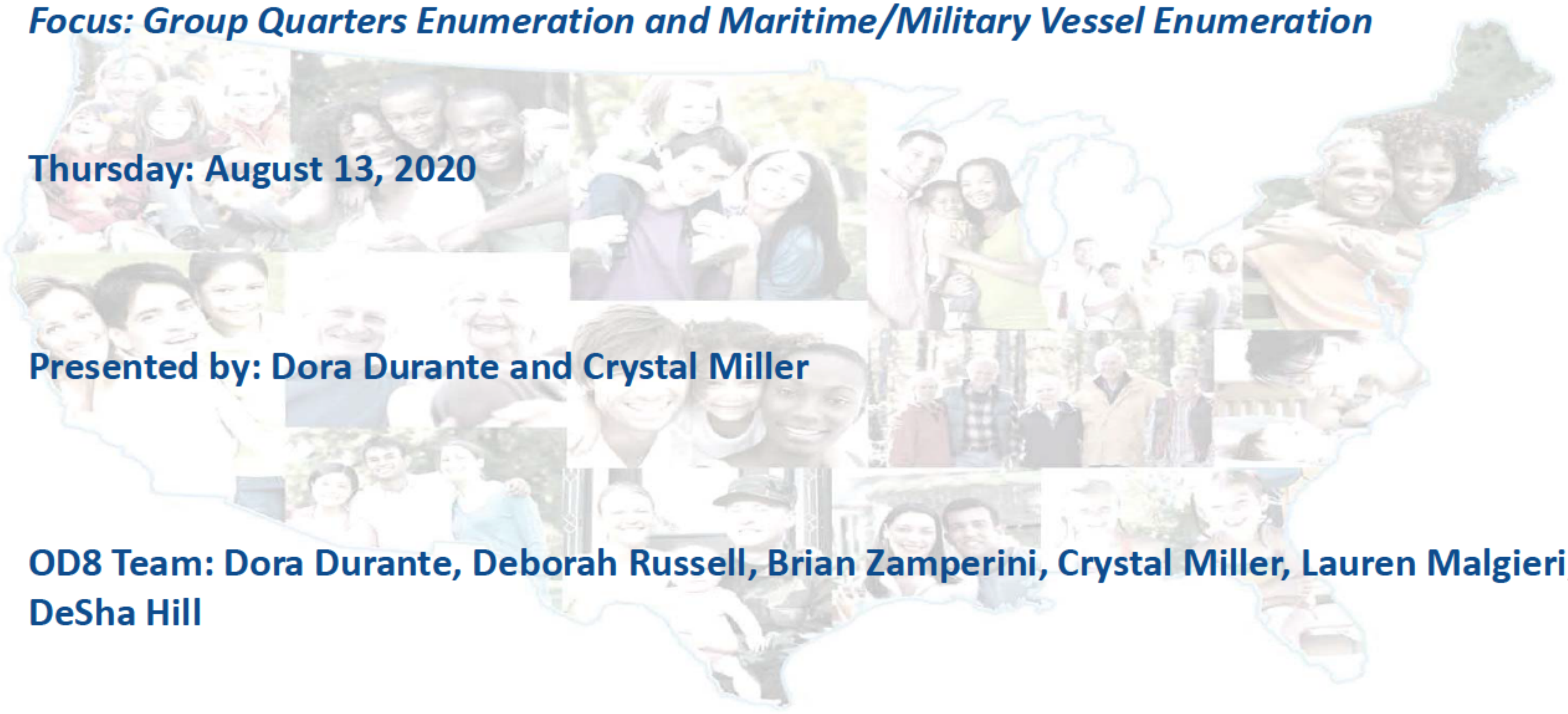
instagram.com/uscensusbureau

Shape your future START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Operational Delivery 8: Group Quarters Update

Focus: Group Quarters Enumeration and Maritime/Military Vessel Enumeration



Thursday: August 13, 2020

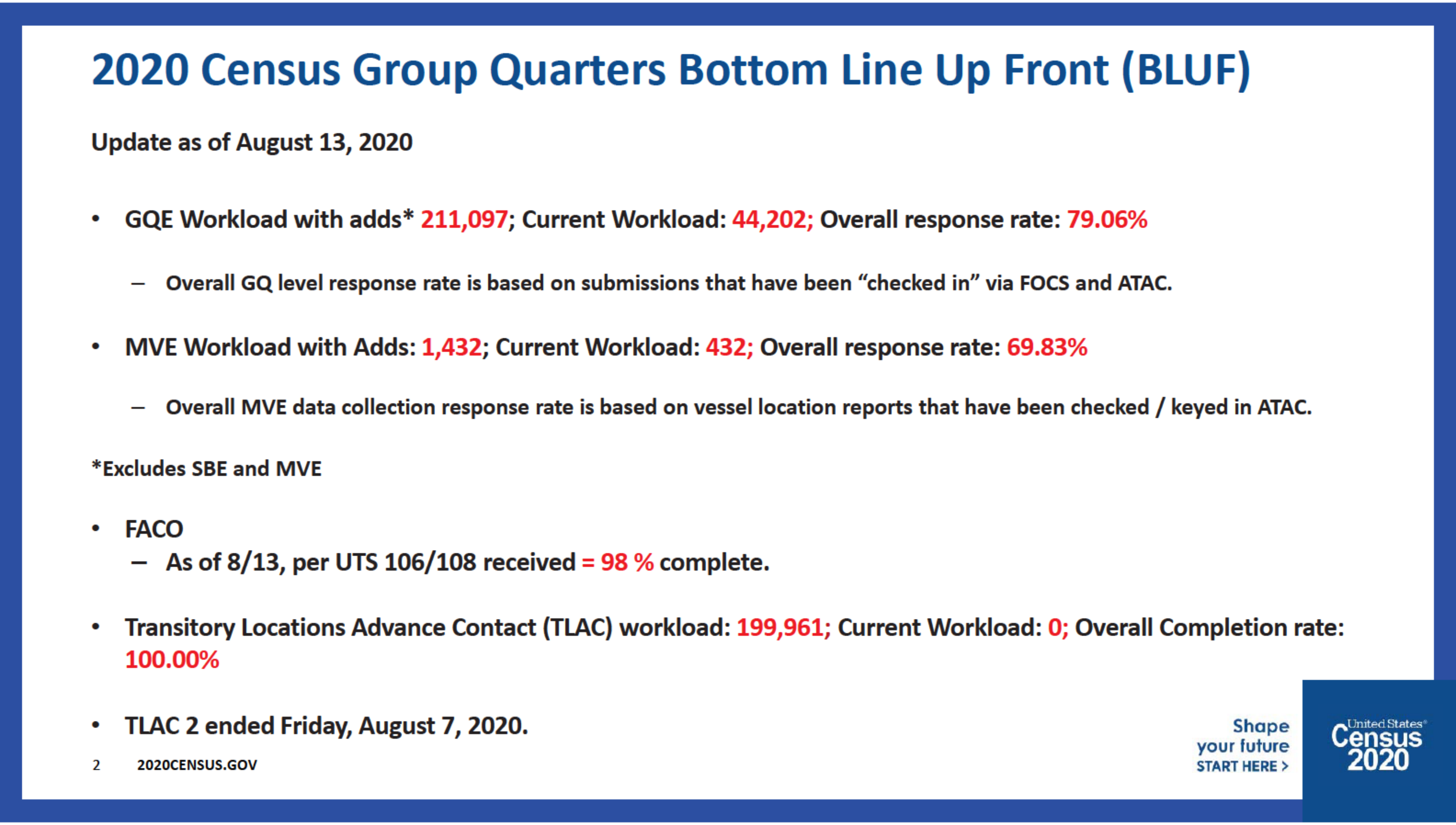
Presented by: Dora Durante and Crystal Miller

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

Shape
your future
START HERE >

United States®
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census Group Quarters Bottom Line Up Front (BLUF)

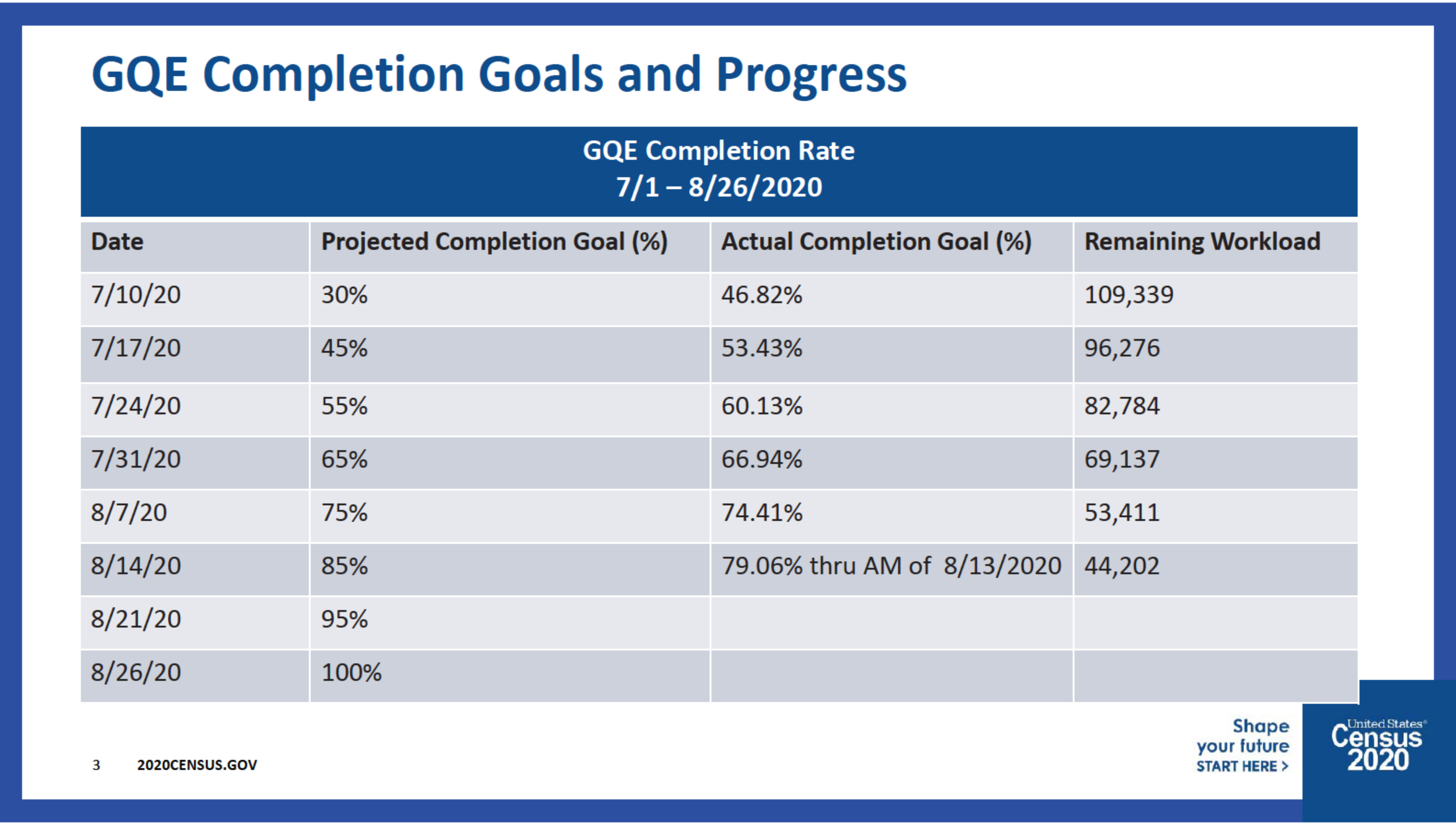
Update as of August 13, 2020

- GQE Workload with adds* 211,097; Current Workload: 44,202; Overall response rate: 79.06%
  - Overall GQ level response rate is based on submissions that have been "checked in" via FOCS and ATAC.
- MVE Workload with Adds: 1,432; Current Workload: 432; Overall response rate: 69.83%
  - Overall MVE data collection response rate is based on vessel location reports that have been checked / keyed in ATAC.

*Excludes SBE and MVE

- FACO
  - As of 8/13, per UTS 106/108 received = 98 % complete.
- Transitory Locations Advance Contact (TLAC) workload: 199,961; Current Workload: 0; Overall Completion rate: 100.00%
- TLAC 2 ended Friday, August 7, 2020.

Shape
your future
START HERE >

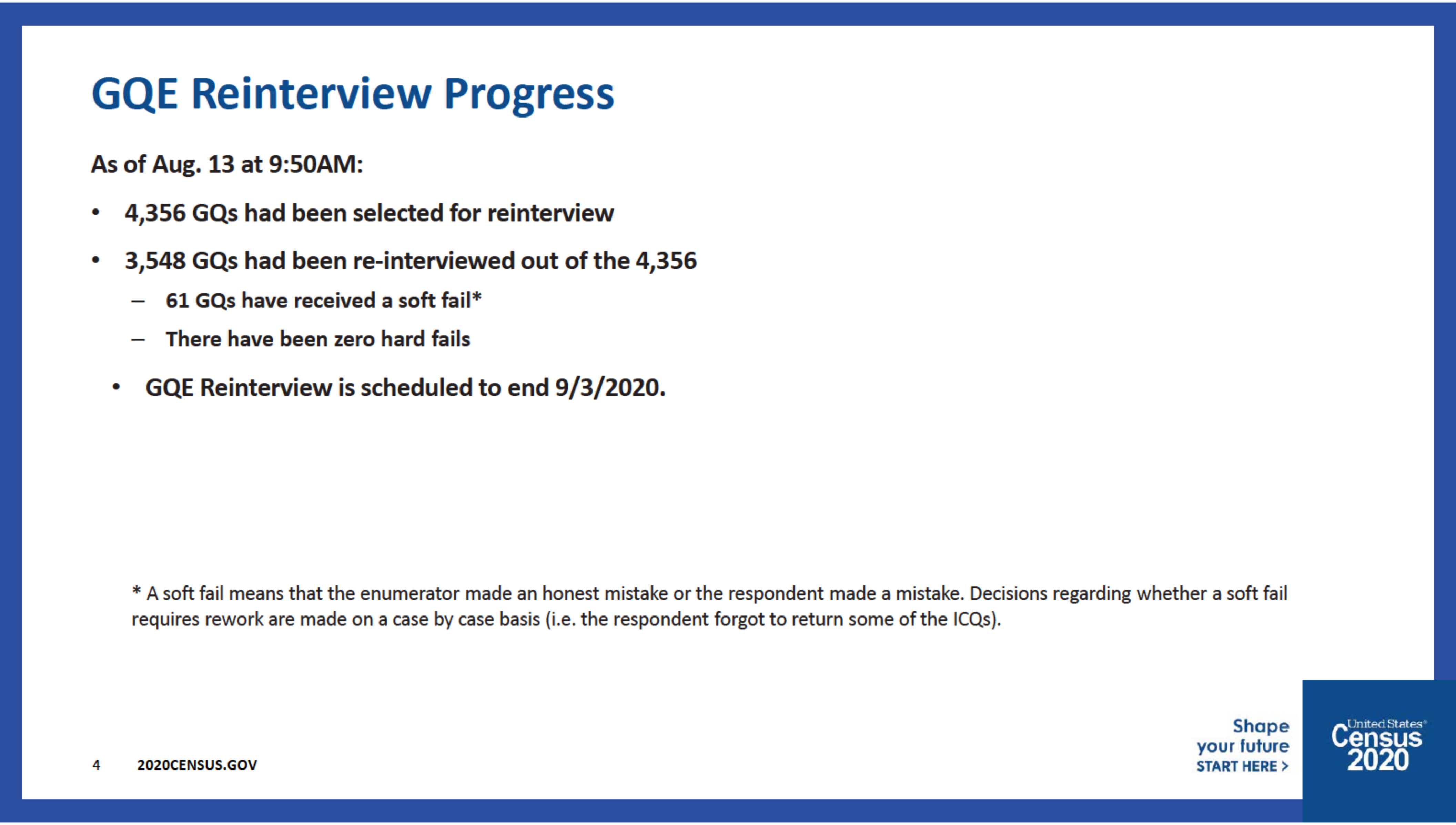


GQE Completion Goals and Progress

GQE Completion Rate 7/1 – 8/26/2020			
Date	**Projected Completion Goal (%)**	**Actual Completion Goal (%)**	**Remaining Workload**
7/10/20	30%	46.82%	109,339
7/17/20	45%	53.43%	96,276
7/24/20	55%	60.13%	82,784
7/31/20	65%	66.94%	69,137
8/7/20	75%	74.41%	53,411
8/14/20	85%	79.06% thru AM of 8/13/2020	44,202
8/21/20	95%		
8/26/20	100%		

Shape your future START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



GQE Reinterview Progress

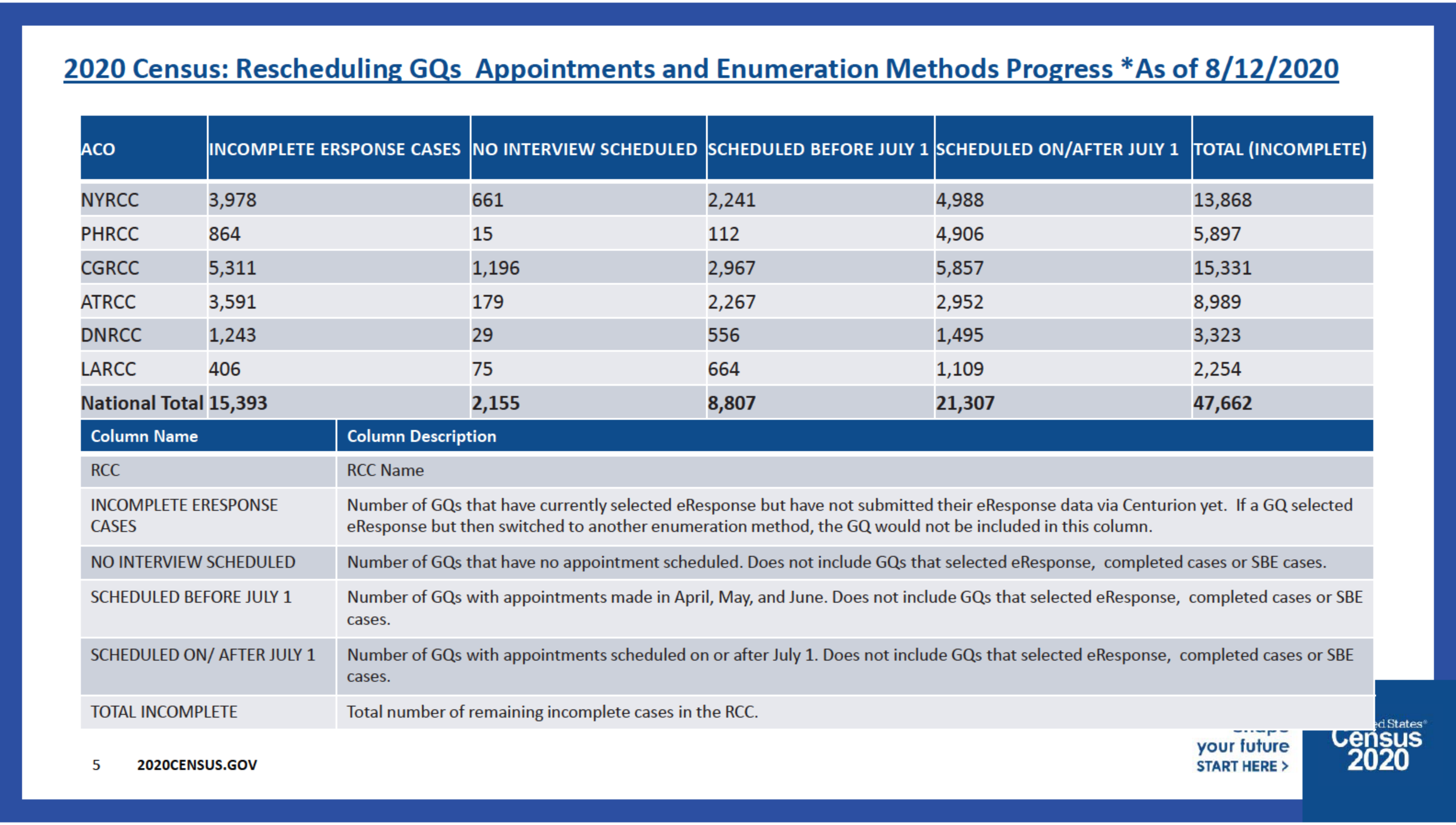
As of Aug. 13 at 9:50AM:

- **4,356 GQs had been selected for reinterview**
- **3,548 GQs had been re-interviewed out of the 4,356**
  - **61 GQs have received a soft fail***
  - **There have been zero hard fails**
- **GQE Reinterview is scheduled to end 9/3/2020.**

* A soft fail means that the enumerator made an honest mistake or the respondent made a mistake. Decisions regarding whether a soft fail requires rework are made on a case by case basis (i.e. the respondent forgot to return some of the ICQs).

Shape your future START HERE >

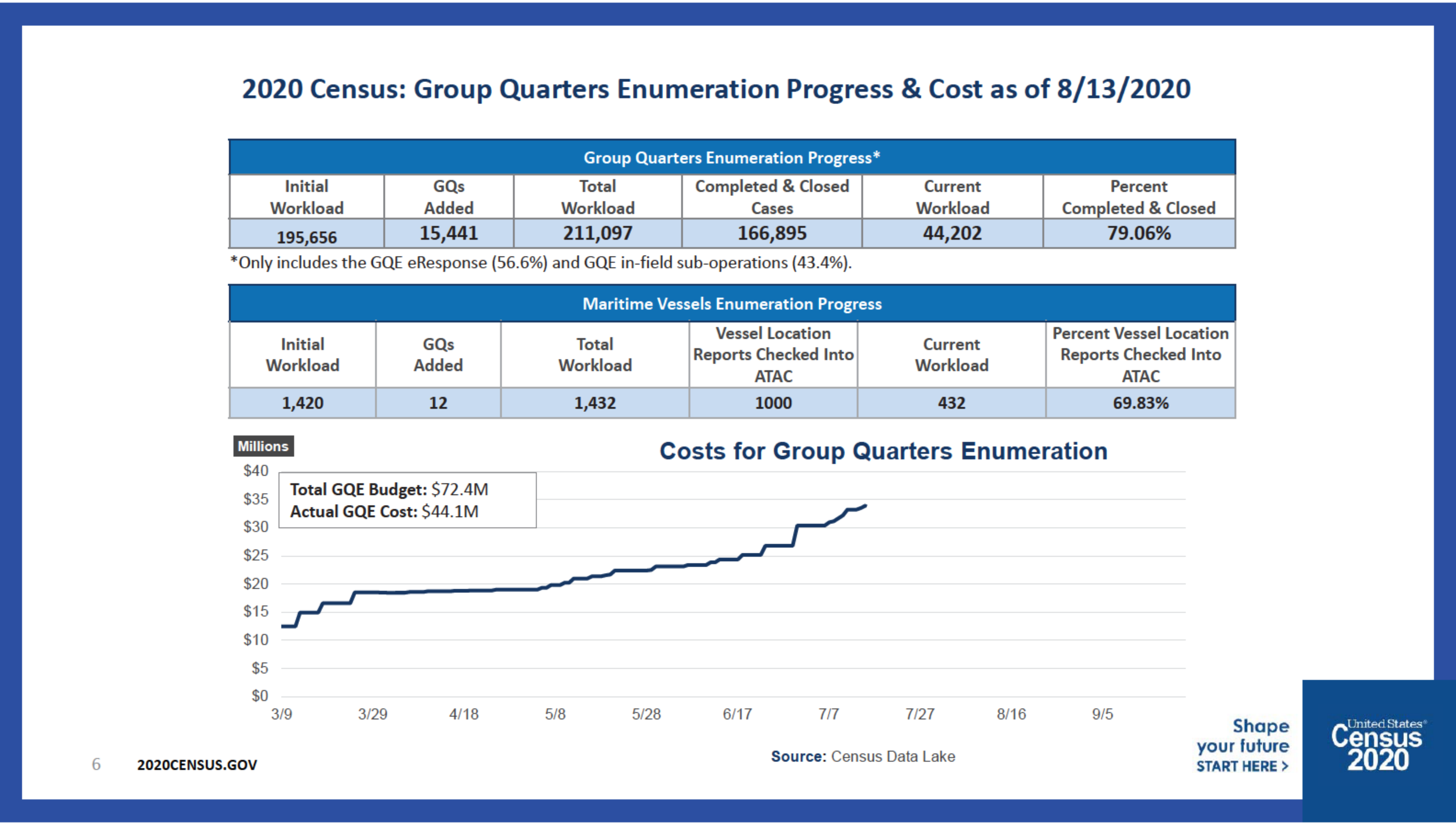
DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census: Rescheduling GQs Appointments and Enumeration Methods Progress *As of 8/12/2020

ACO	INCOMPLETE ERSPONSE CASES	NO INTERVIEW SCHEDULED	SCHEDULED BEFORE JULY 1	SCHEDULED ON/AFTER JULY 1	TOTAL (INCOMPLETE)
NYRCC	3,978	661	2,241	4,988	13,868
PHRCC	864	15	112	4,906	5,897
CGRCC	5,311	1,196	2,967	5,857	15,331
ATRCC	3,591	179	2,267	2,952	8,989
DNRCC	1,243	29	556	1,495	3,323
LARCC	406	75	664	1,109	2,254
National Total	**15,393**	**2,155**	**8,807**	**21,307**	**47,662**

Column Name	Column Description
RCC	RCC Name
INCOMPLETE ERESPONSE CASES	Number of GQs that have currently selected eResponse but have not submitted their eResponse data via Centurion yet. If a GQ selected eResponse but then switched to another enumeration method, the GQ would not be included in this column.
NO INTERVIEW SCHEDULED	Number of GQs that have no appointment scheduled. Does not include GQs that selected eResponse, completed cases or SBE cases.
SCHEDULED BEFORE JULY 1	Number of GQs with appointments made in April, May, and June. Does not include GQs that selected eResponse, completed cases or SBE cases.
SCHEDULED ON/ AFTER JULY 1	Number of GQs with appointments scheduled on or after July 1. Does not include GQs that selected eResponse, completed cases or SBE cases.
TOTAL INCOMPLETE	Total number of remaining incomplete cases in the RCC.



2020 Census: Group Quarters Enumeration Progress & Cost as of 8/13/2020

Group Quarters Enumeration Progress*					
Initial Workload	GQs Added	Total Workload	Completed & Closed Cases	Current Workload	Percent Completed & Closed
195,656	15,441	211,097	166,895	44,202	79.06%

*Only includes the GQE eResponse (56.6%) and GQE in-field sub-operations (43.4%).

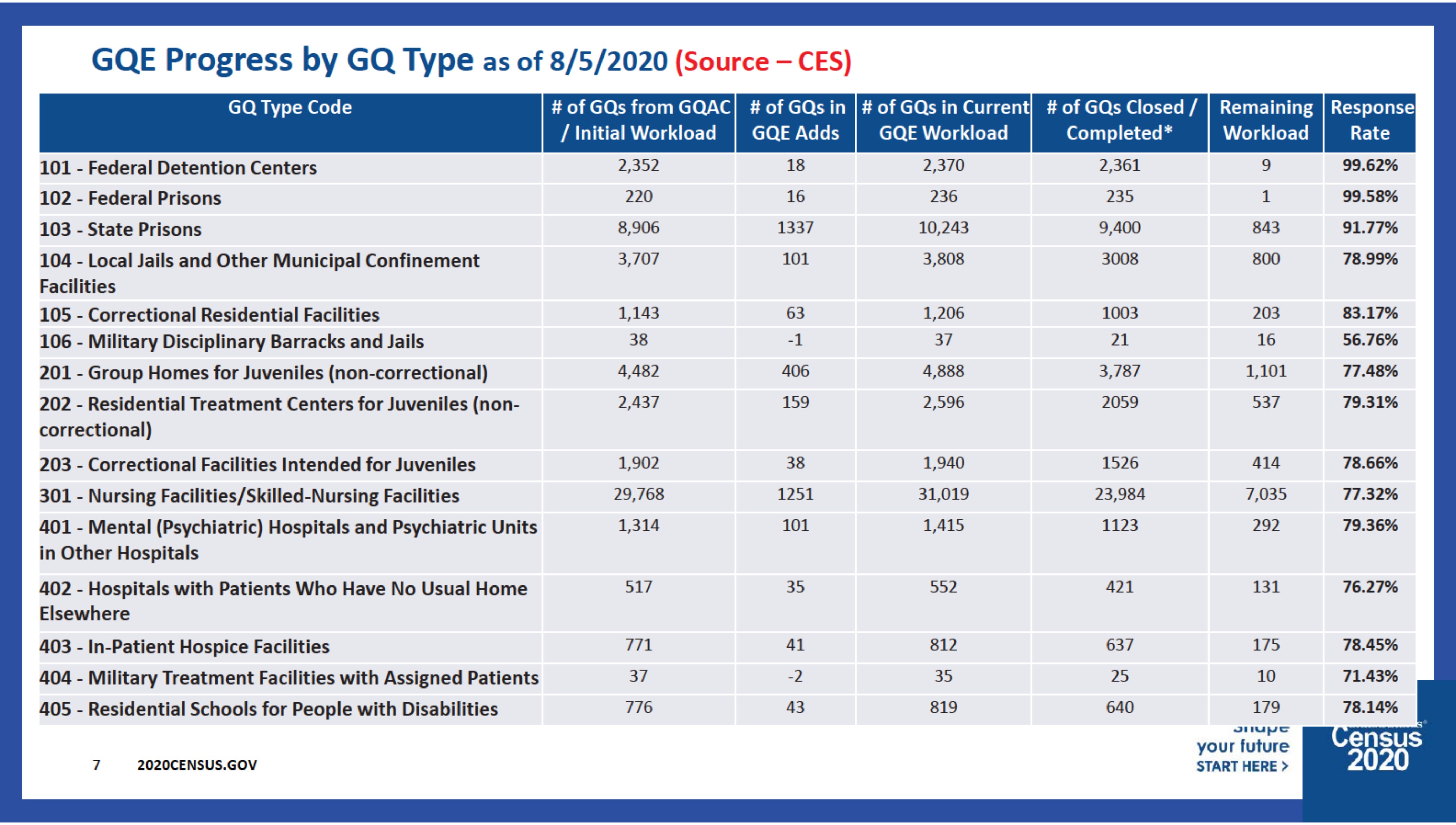
Maritime Vessels Enumeration Progress					
Initial Workload	GQs Added	Total Workload	Vessel Location Reports Checked Into ATAC	Current Workload	Percent Vessel Location Reports Checked Into ATAC
1,420	12	1,432	1000	432	69.83%

Costs for Group Quarters Enumeration

Total GQE Budget: $72.4M
Actual GQE Cost: $44.1M

Source: Census Data Lake

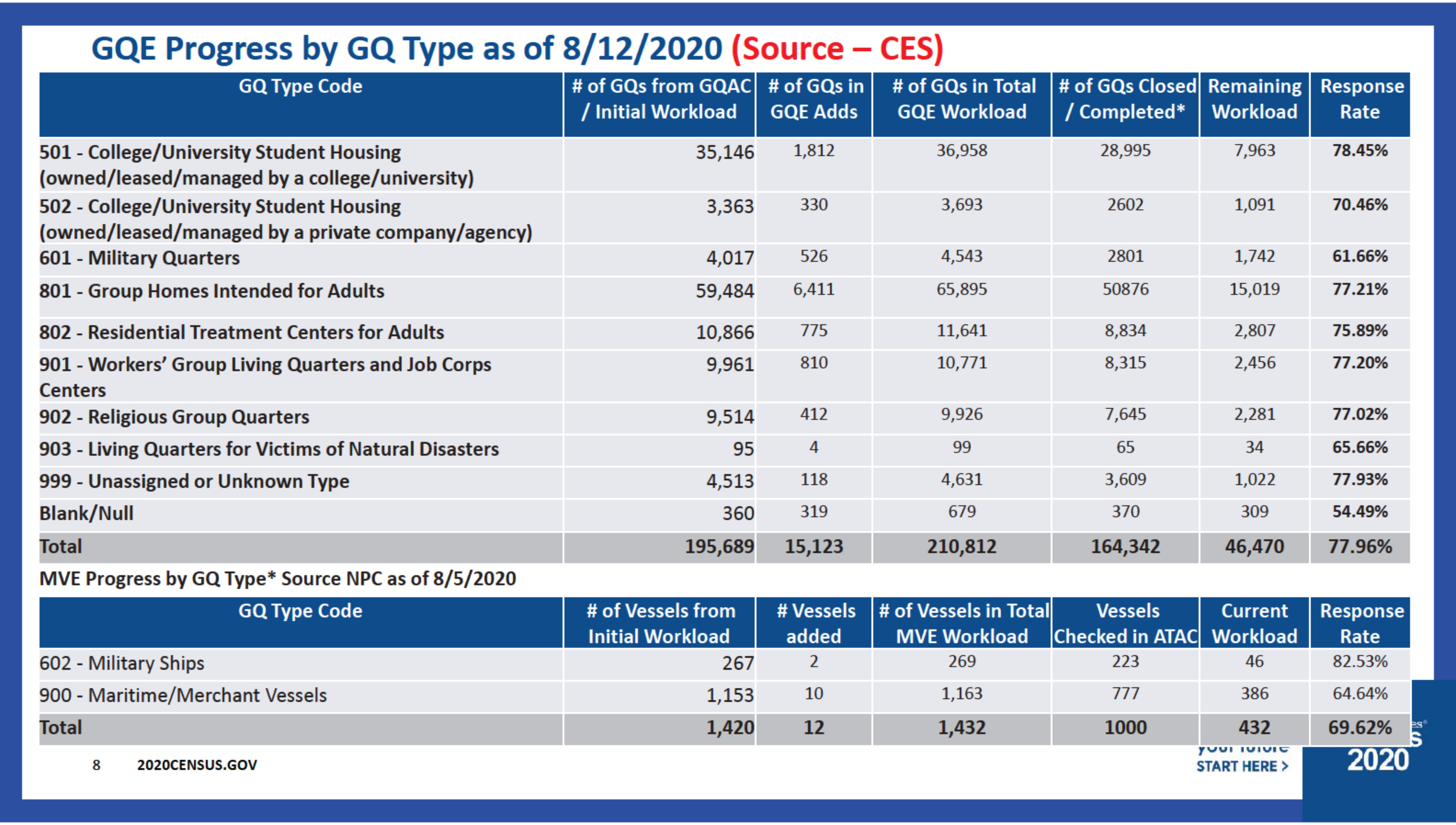
DRB Approval Number: CBDRB-FY21-DSEP-002



GQE Progress by GQ Type as of 8/5/2020 (Source – CES)

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Current GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
101 - Federal Detention Centers	2,352	18	2,370	2,361	9	**99.62%**
102 - Federal Prisons	220	16	236	235	1	**99.58%**
103 - State Prisons	8,906	1337	10,243	9,400	843	**91.77%**
104 - Local Jails and Other Municipal Confinement Facilities	3,707	101	3,808	3008	800	**78.99%**
105 - Correctional Residential Facilities	1,143	63	1,206	1003	203	**83.17%**
106 - Military Disciplinary Barracks and Jails	38	-1	37	21	16	**56.76%**
201 - Group Homes for Juveniles (non-correctional)	4,482	406	4,888	3,787	1,101	**77.48%**
202 - Residential Treatment Centers for Juveniles (non-correctional)	2,437	159	2,596	2059	537	**79.31%**
203 - Correctional Facilities Intended for Juveniles	1,902	38	1,940	1526	414	**78.66%**
301 - Nursing Facilities/Skilled-Nursing Facilities	29,768	1251	31,019	23,984	7,035	**77.32%**
401 - Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals	1,314	101	1,415	1123	292	**79.36%**
402 - Hospitals with Patients Who Have No Usual Home Elsewhere	517	35	552	421	131	**76.27%**
403 - In-Patient Hospice Facilities	771	41	812	637	175	**78.45%**
404 - Military Treatment Facilities with Assigned Patients	37	-2	35	25	10	**71.43%**
405 - Residential Schools for People with Disabilities	776	43	819	640	179	**78.14%**

DRB Approval Number: CBDRB-FY21-DSEP-002

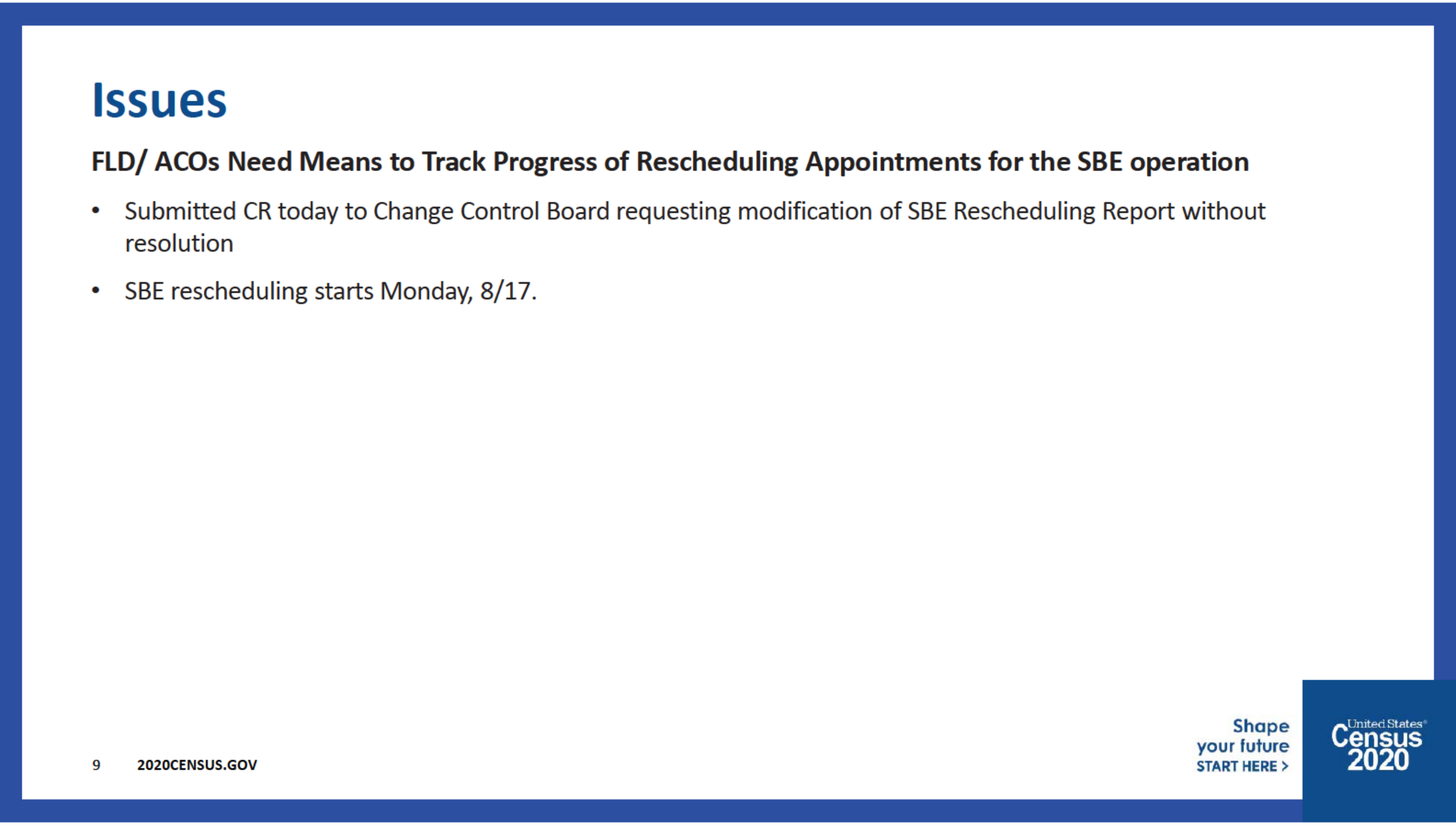


GQE Progress by GQ Type as of 8/12/2020 (Source – CES)

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Total GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
501 - College/University Student Housing (owned/leased/managed by a college/university)	35,146	1,812	36,958	28,995	7,963	**78.45%**
502 - College/University Student Housing (owned/leased/managed by a private company/agency)	3,363	330	3,693	2602	1,091	**70.46%**
601 - Military Quarters	4,017	526	4,543	2801	1,742	**61.66%**
801 - Group Homes Intended for Adults	59,484	6,411	65,895	50876	15,019	**77.21%**
802 - Residential Treatment Centers for Adults	10,866	775	11,641	8,834	2,807	**75.89%**
901 - Workers' Group Living Quarters and Job Corps Centers	9,961	810	10,771	8,315	2,456	**77.20%**
902 - Religious Group Quarters	9,514	412	9,926	7,645	2,281	**77.02%**
903 - Living Quarters for Victims of Natural Disasters	95	4	99	65	34	**65.66%**
999 - Unassigned or Unknown Type	4,513	118	4,631	3,609	1,022	**77.93%**
Blank/Null	360	319	679	370	309	**54.49%**
Total	**195,689**	**15,123**	**210,812**	**164,342**	**46,470**	**77.96%**

MVE Progress by GQ Type* Source NPC as of 8/5/2020

GQ Type Code	# of Vessels from Initial Workload	# Vessels added	# of Vessels in Total MVE Workload	Vessels Checked in ATAC	Current Workload	Response Rate
602 - Military Ships	267	2	269	223	46	82.53%
900 - Maritime/Merchant Vessels	1,153	10	1,163	777	386	64.64%
Total	**1,420**	**12**	**1,432**	**1000**	**432**	**69.62%**



Issues

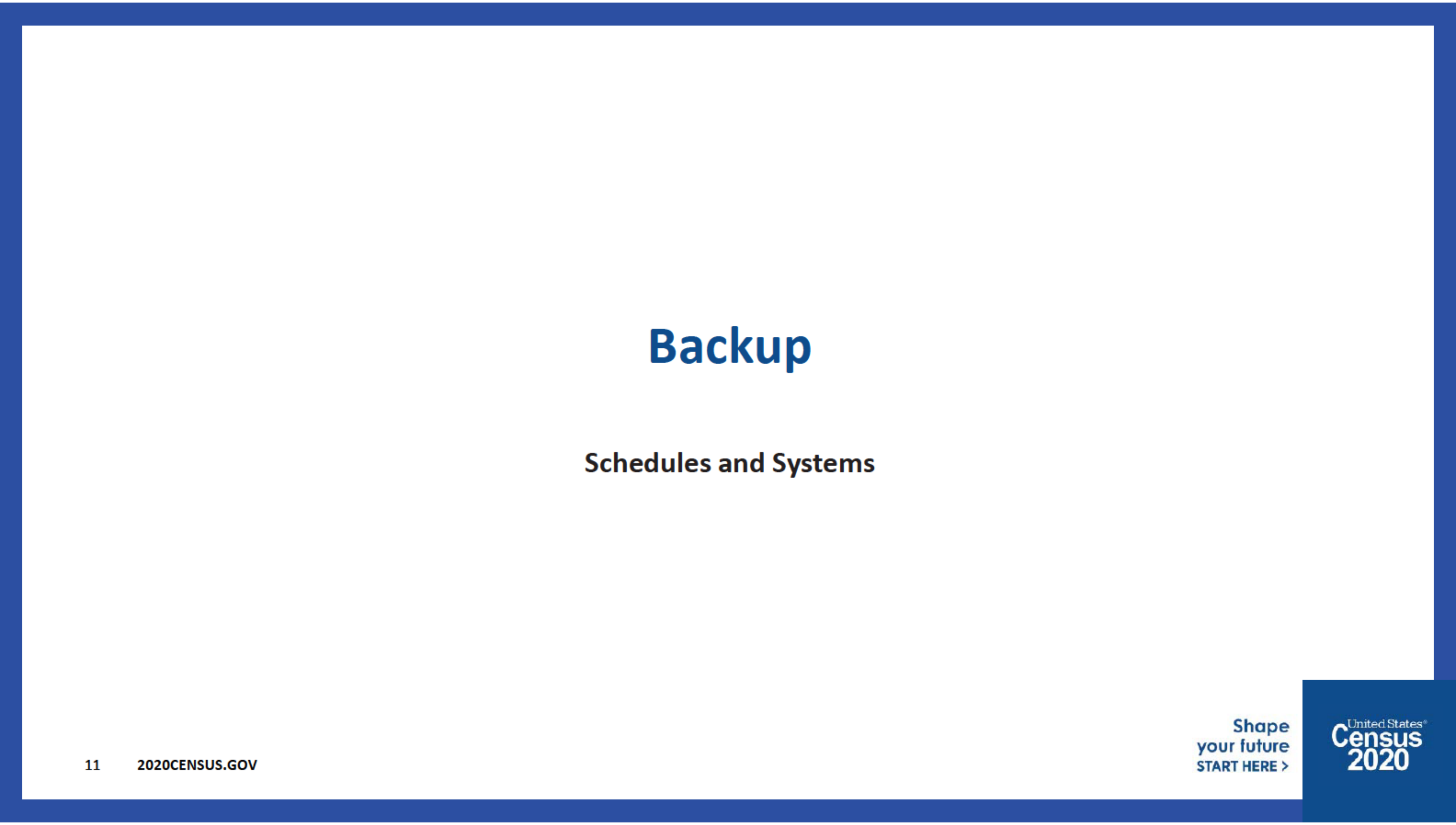
FLD/ ACOs Need Means to Track Progress of Rescheduling Appointments for the SBE operation

- Submitted CR today to Change Control Board requesting modification of SBE Rescheduling Report without resolution
- SBE rescheduling starts Monday, 8/17.

Apportionment Timeline Impacts

- **Late GQE -** Canceled
- **GQE and GQ Maritime -** Working with stakeholders to ensure all dependencies are met to make the final collection universe and CUF processing (e.g. ADDUPs, Response Processing)
- **ETL and SBE -** Discussions ongoing for the following
  - Obtaining POP counts by State for inclusion in Apportionment and ensuring this matches counts in Redistricting
  - Including response data in Redistricting that is geocoded to the block level
- **FACO/Deployed Military –** Working to update the schedule/runbook to meet the FACO Final Delivery date including the unmatched deployed file

DRB Approval Number: CBDRB-FY21-DSEP-002



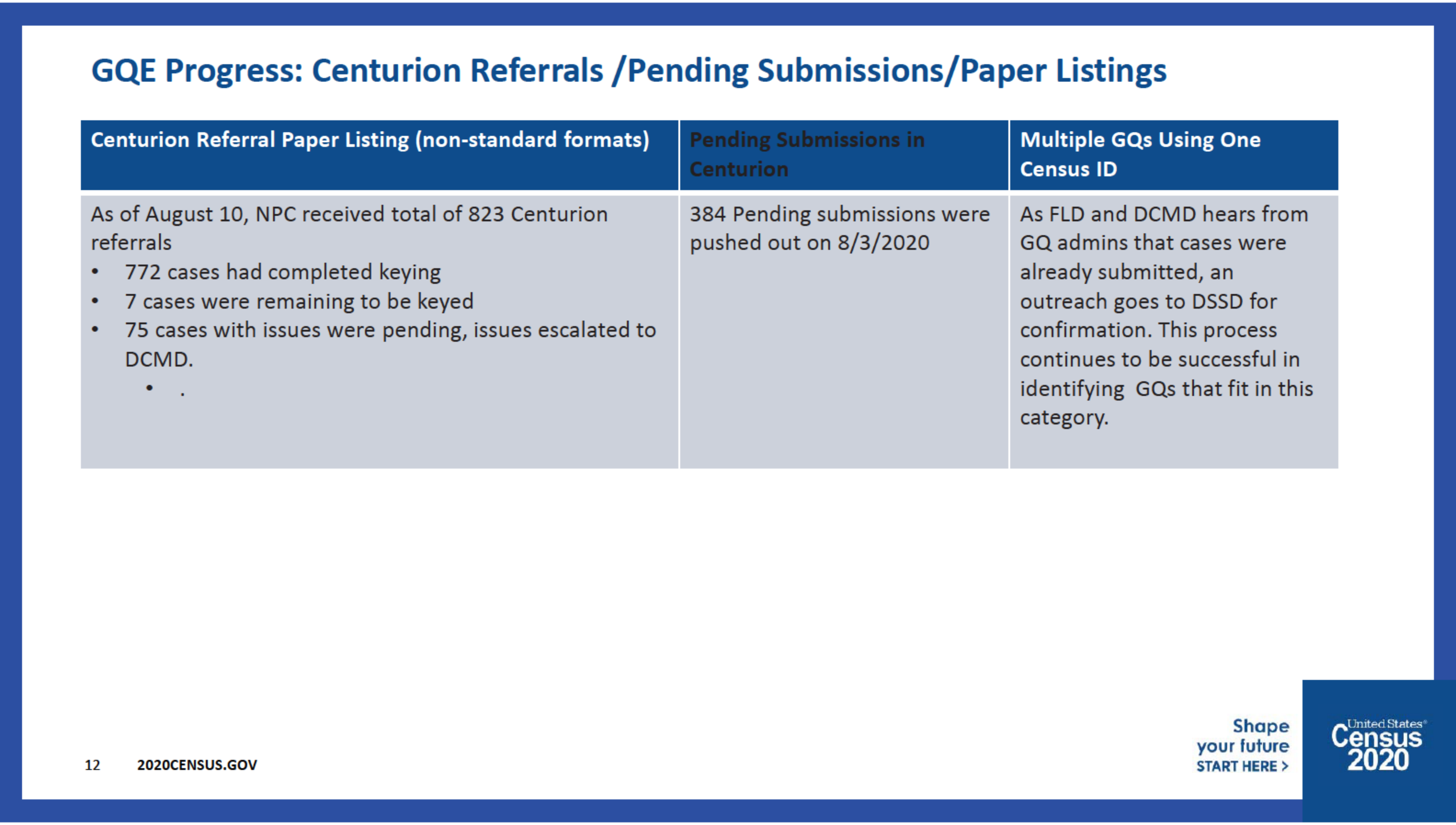
Backup

Schedules and Systems

Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

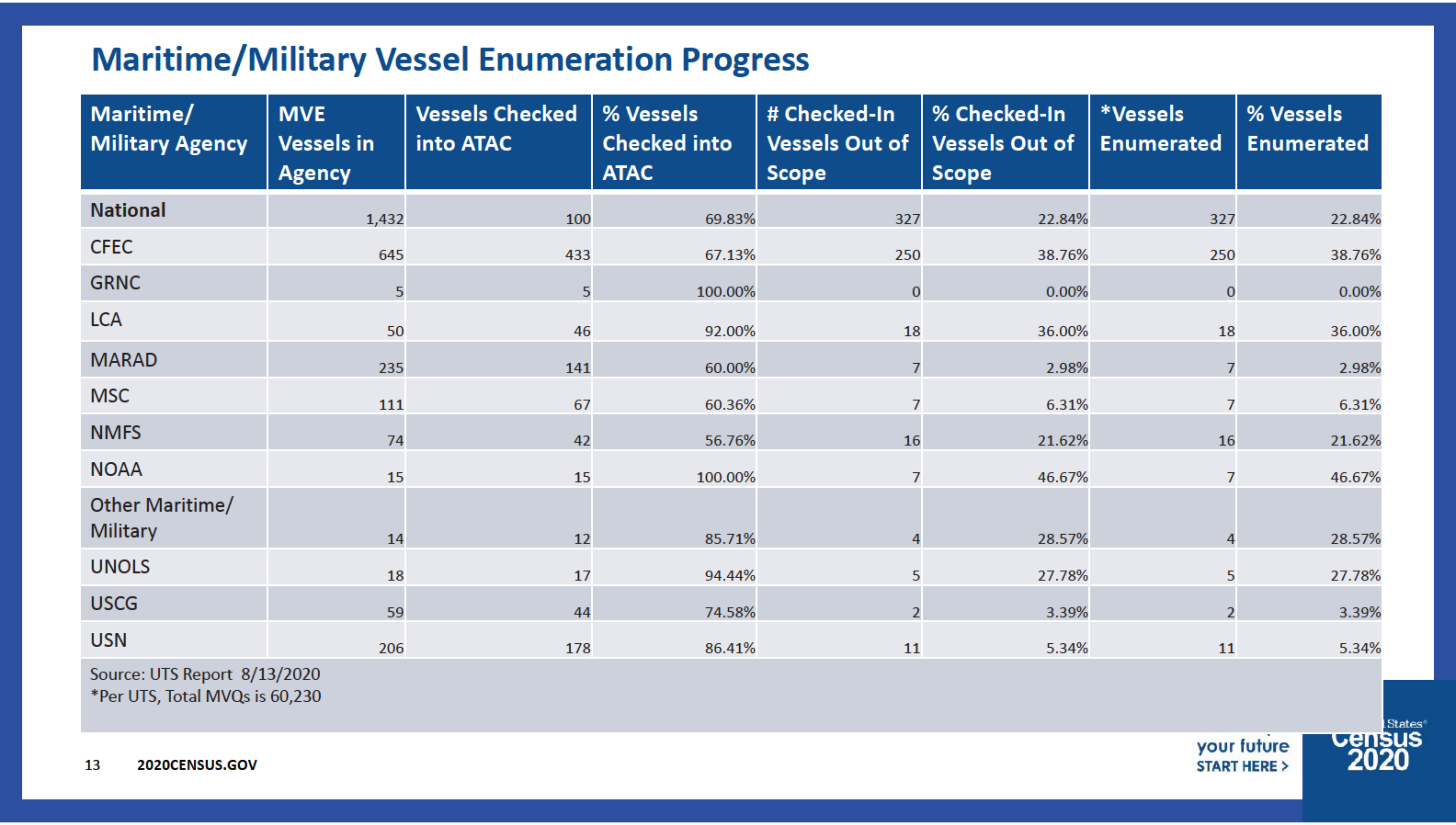


GQE Progress: Centurion Referrals /Pending Submissions/Paper Listings

Centurion Referral Paper Listing (non-standard formats)	Pending Submissions in Centurion	Multiple GQs Using One Census ID
As of August 10, NPC received total of 823 Centurion referrals • 772 cases had completed keying • 7 cases were remaining to be keyed • 75 cases with issues were pending, issues escalated to DCMD. • .	384 Pending submissions were pushed out on 8/3/2020	As FLD and DCMD hears from GQ admins that cases were already submitted, an outreach goes to DSSD for confirmation. This process continues to be successful in identifying GQs that fit in this category.

Shape
your future
START HERE >

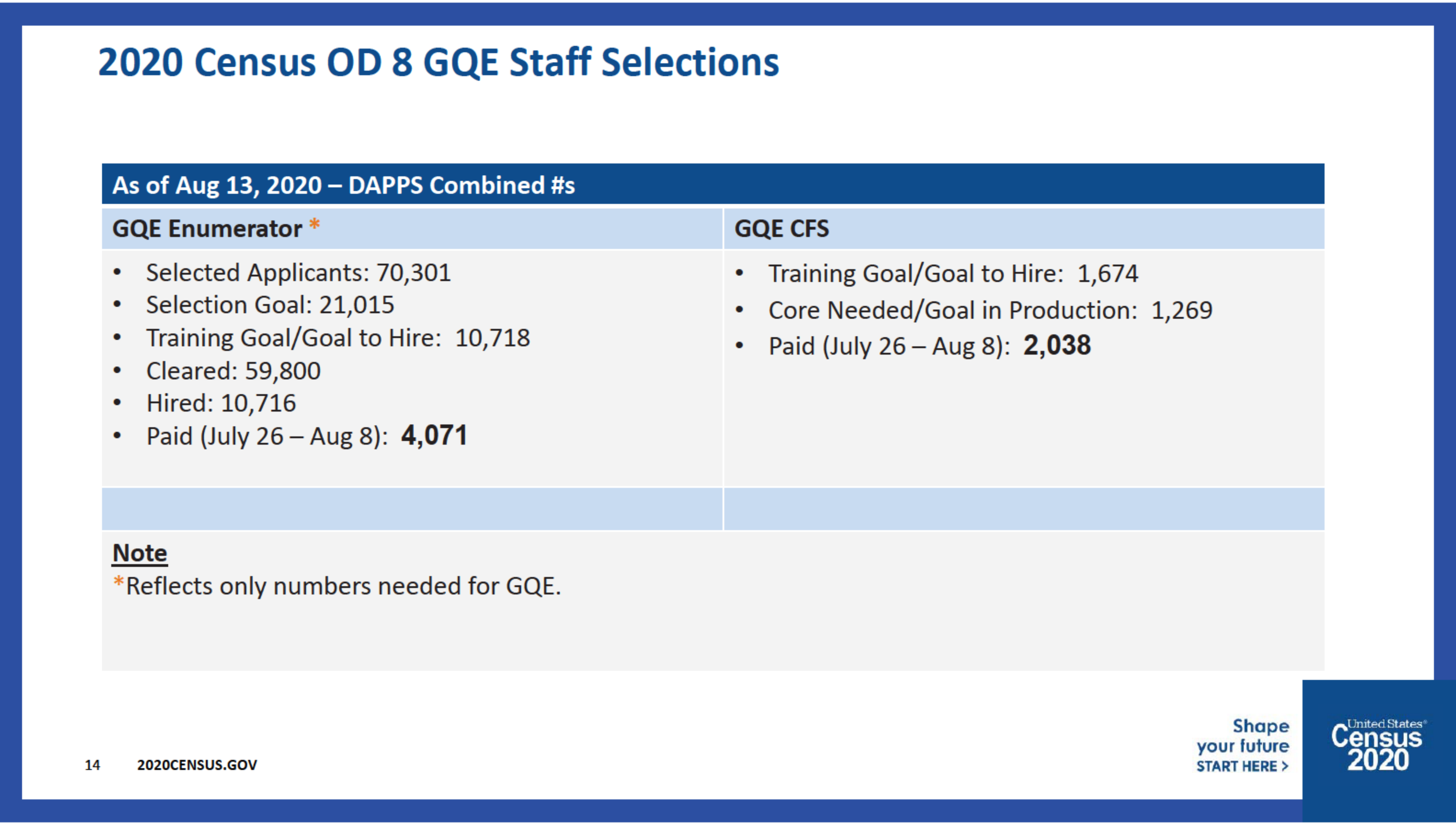
DRB Approval Number: CBDRB-FY21-DSEP-002



Maritime/Military Vessel Enumeration Progress

Maritime/ Military Agency	MVE Vessels in Agency	Vessels Checked into ATAC	% Vessels Checked into ATAC	# Checked-In Vessels Out of Scope	% Checked-In Vessels Out of Scope	*Vessels Enumerated	% Vessels Enumerated
National	1,432	100	69.83%	327	22.84%	327	22.84%
CFEC	645	433	67.13%	250	38.76%	250	38.76%
GRNC	5	5	100.00%	0	0.00%	0	0.00%
LCA	50	46	92.00%	18	36.00%	18	36.00%
MARAD	235	141	60.00%	7	2.98%	7	2.98%
MSC	111	67	60.36%	7	6.31%	7	6.31%
NMFS	74	42	56.76%	16	21.62%	16	21.62%
NOAA	15	15	100.00%	7	46.67%	7	46.67%
Other Maritime/ Military	14	12	85.71%	4	28.57%	4	28.57%
UNOLS	18	17	94.44%	5	27.78%	5	27.78%
USCG	59	44	74.58%	2	3.39%	2	3.39%
USN	206	178	86.41%	11	5.34%	11	5.34%

Source: UTS Report 8/13/2020
*Per UTS, Total MVQs is 60,230



2020 Census OD 8 GQE Staff Selections

As of Aug 13, 2020 – DAPPS Combined #s

GQE Enumerator *	GQE CFS
• Selected Applicants: 70,301 • Selection Goal: 21,015 • Training Goal/Goal to Hire: 10,718 • Cleared: 59,800 • Hired: 10,716 • Paid (July 26 – Aug 8): **4,071**	• Training Goal/Goal to Hire: 1,674 • Core Needed/Goal in Production: 1,269 • Paid (July 26 – Aug 8): **2,038**

Note

*Reflects only numbers needed for GQE.

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

OD 8 GQE Milestone Conduct & Training Dates

Conduct Activity	Proposed Start	Proposed Finish
Conduct MVE Operation	04/01/20 (A)	07/31/20 (P)
Conduct GQE eResponse Operation	04/01/20 (A)	08/07/20 (P)
Conduct GQE Operation Field	04/20/20 (A)	08/26/20 (P)
Conduct GQE In-Person Operation	07/01/20 (A)	08/26/20 (P)
Conduct GQ Reinterview Operation	07/02/20 (A)	09/03/20 (P)
Conduct DVS Enumeration Field	07/06/20 (A)	08/26/20 (P)

OD 8 GQE, SBE, ETL, MVE & Military Operational Outreach Throughout COVID

The GQO team has been assisting GQ Admins (Gatekeepers) with their 2020 Census submission by performing the following tasks:

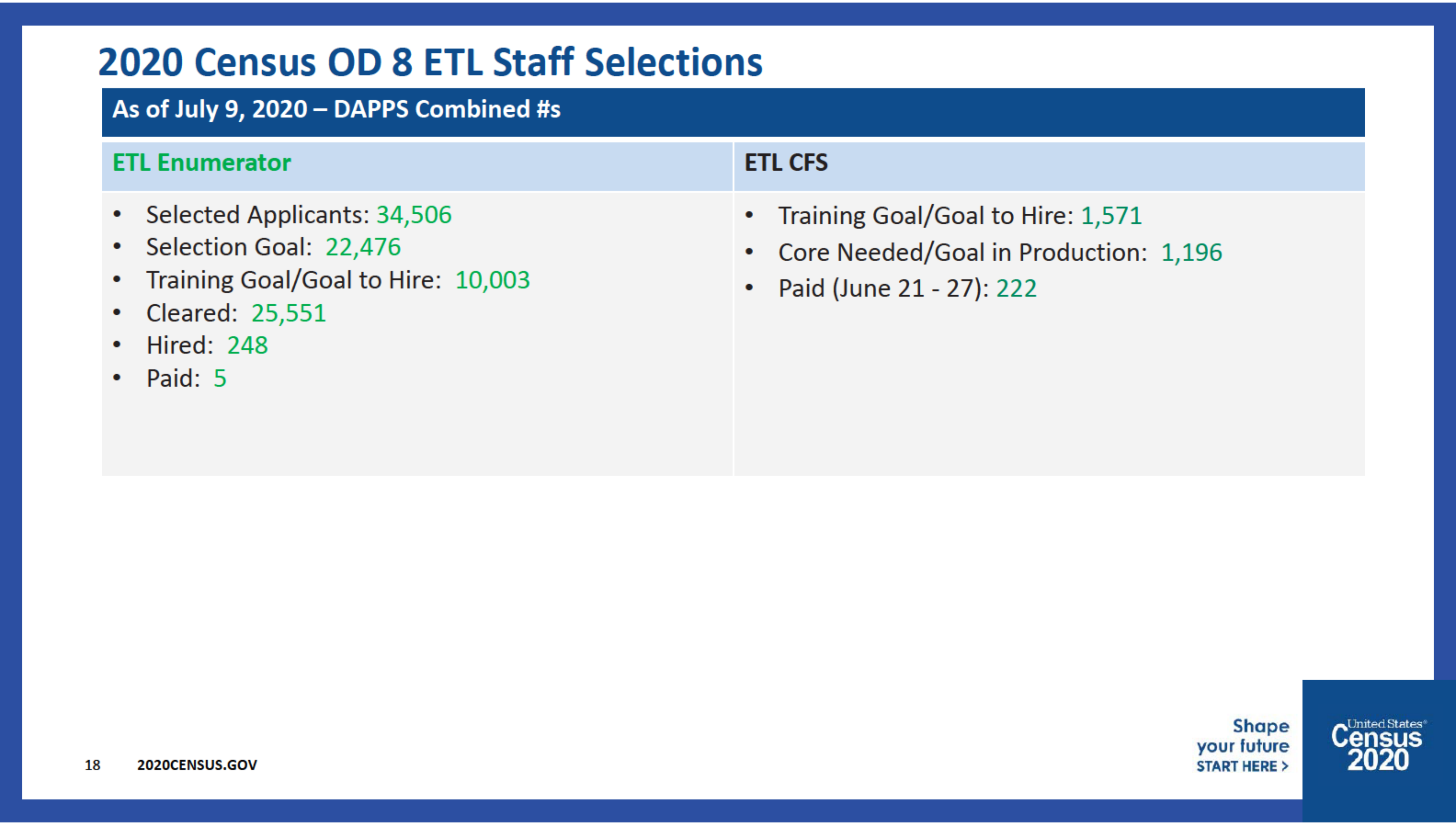
- Responding to emails and phone calls from GQ administrators / ACO staff
- Transcribing data from Paper Listings/ non-standard formats into eResponse standard template
- Uploading eResponse submittals and walking GQ administrators through submissions
- Working with Legal to create letters to get refusing GQ administrators to respond
- Scheduling and participating in meetings with refusing GQ administrators to encourage participation and responses

2020 eResponse Helpdesk Weekly Update (8/7-8/12)	
Number of staff/volunteers working	10
Total Numbers of Hours Spent	210
Average number of emails received/responded	150
Average number of telephone calls/Voicemails	65
Average Emails/Calls/Cases Resolved	71
Total Uploads completed	35
Cases Referred to FLD and SEB	40
**. Table will be updated each week to show weekly progress.	

DRB Approval Number: CBDRB-FY21-DSEP-002

OD 8 GQE, SBE, ETL, MVE & Military System Updates

- **Box Check In Issue Resolution**
  - Box Check In working group stood up with members from NPC, FOCS, CDL, UTS, DCMD, FLD, TI and DSSD to address box check in issues (which prevent cases from reaching completion status) and drive them to closure
- **Field OCS Updates**
  - CR 1853 approved which allows for more flexibility to allow cases that are currently stuck in a certain status due to user error to move forward
- **FACO**
  - As of 8/13 per UTS 106/108 received = 98% complete.
- **Military**
  - MOB continues to work with the military reps from the CJSWG to get POC updates and resolve issues with base access.



2020 Census OD 8 ETL Staff Selections

As of July 9, 2020 – DAPPS Combined #s

ETL Enumerator

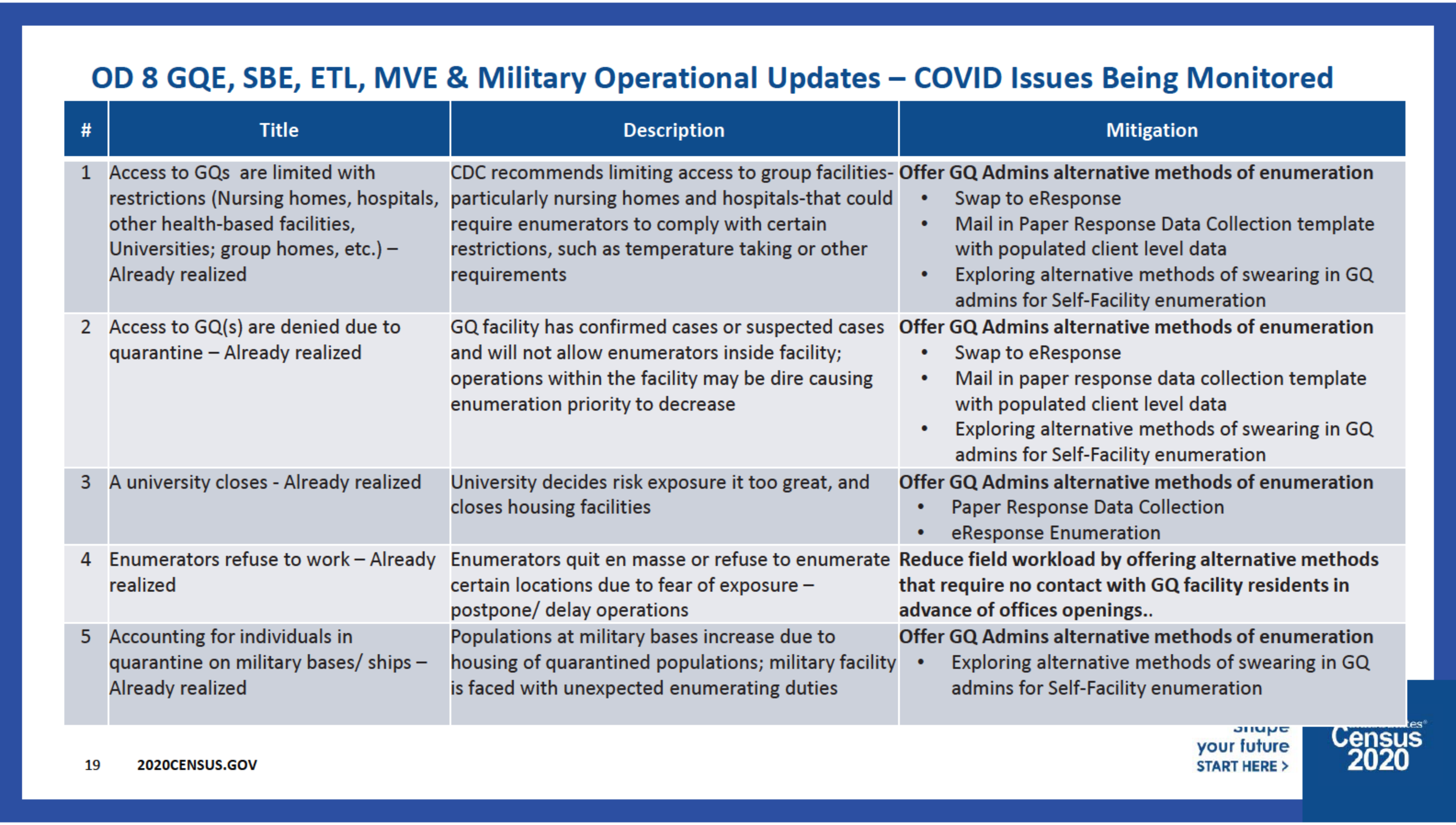
- Selected Applicants: 34,506
- Selection Goal: 22,476
- Training Goal/Goal to Hire: 10,003
- Cleared: 25,551
- Hired: 248
- Paid: 5

ETL CFS

- Training Goal/Goal to Hire: 1,571
- Core Needed/Goal in Production: 1,196
- Paid (June 21 - 27): 222

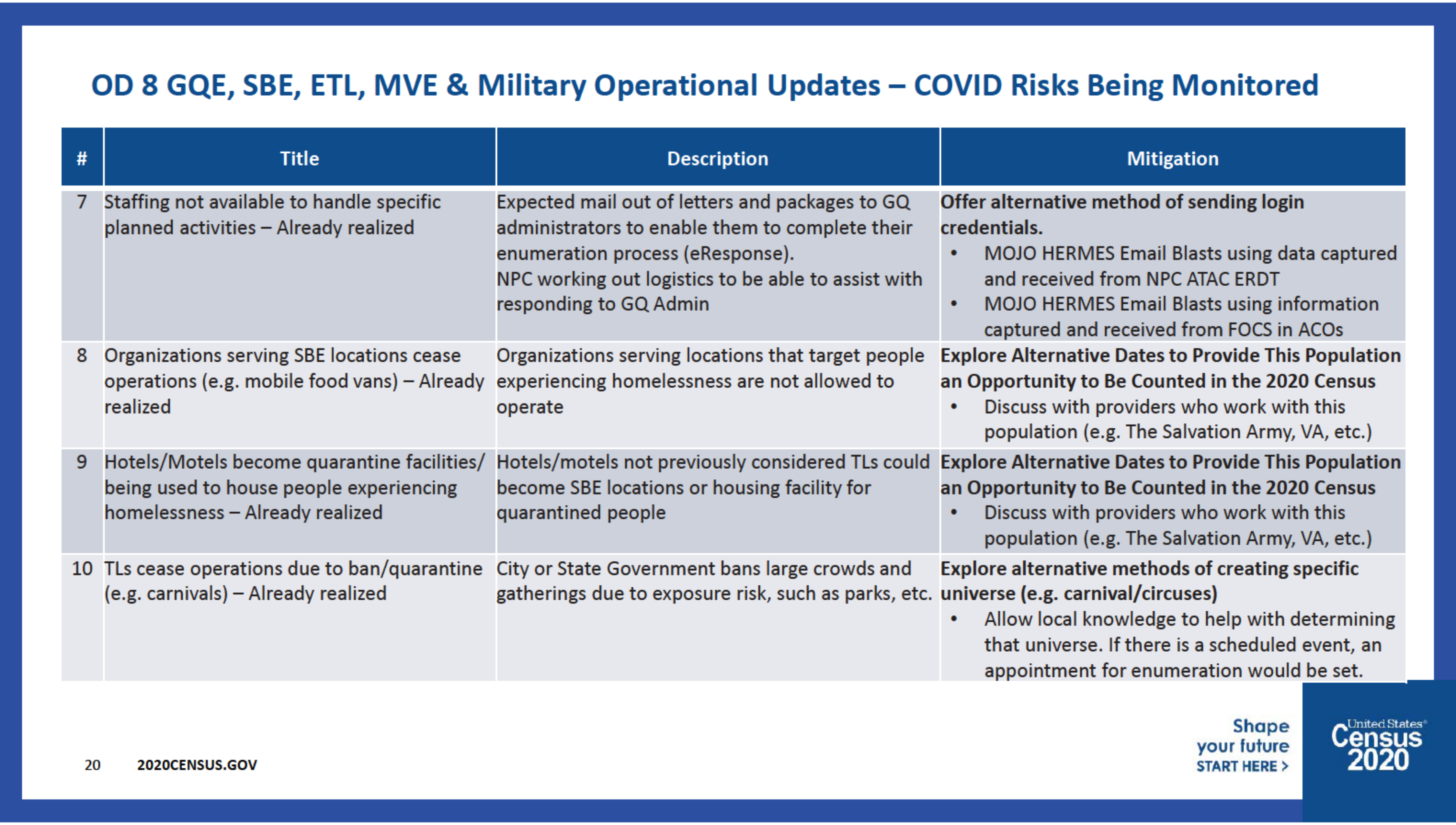
Shape your future START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



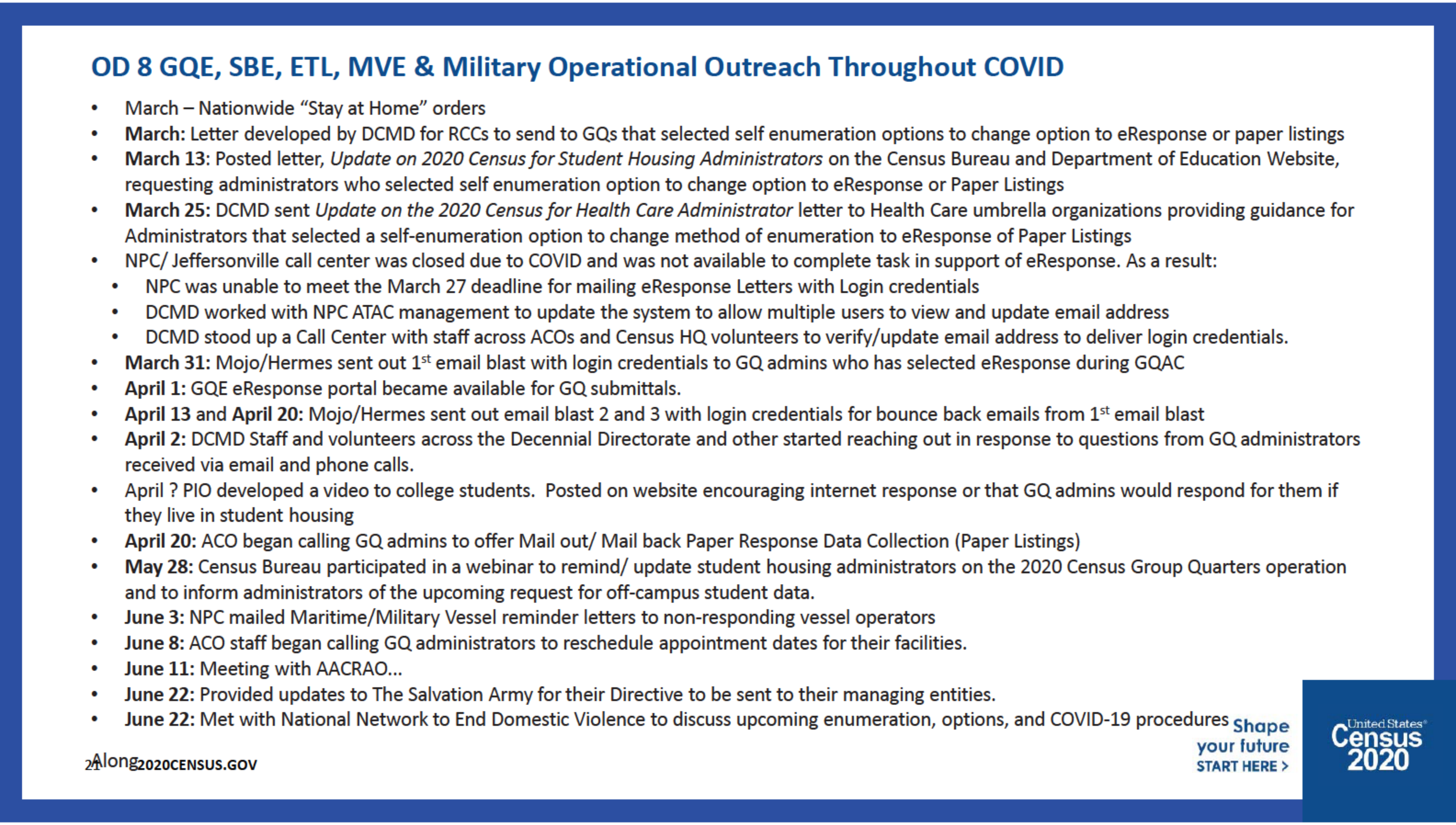
OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Issues Being Monitored

#	Title	Description	Mitigation
1	Access to GQs are limited with restrictions (Nursing homes, hospitals, other health-based facilities, Universities; group homes, etc.) – Already realized	CDC recommends limiting access to group facilities-particularly nursing homes and hospitals-that could require enumerators to comply with certain restrictions, such as temperature taking or other requirements	**Offer GQ Admins alternative methods of enumeration** • Swap to eResponse • Mail in Paper Response Data Collection template with populated client level data • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration
2	Access to GQ(s) are denied due to quarantine – Already realized	GQ facility has confirmed cases or suspected cases and will not allow enumerators inside facility; operations within the facility may be dire causing enumeration priority to decrease	**Offer GQ Admins alternative methods of enumeration** • Swap to eResponse • Mail in paper response data collection template with populated client level data • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration
3	A university closes - Already realized	University decides risk exposure it too great, and closes housing facilities	**Offer GQ Admins alternative methods of enumeration** • Paper Response Data Collection • eResponse Enumeration
4	Enumerators refuse to work – Already realized	Enumerators quit en masse or refuse to enumerate certain locations due to fear of exposure – postpone/ delay operations	**Reduce field workload by offering alternative methods that require no contact with GQ facility residents in advance of offices openings..**
5	Accounting for individuals in quarantine on military bases/ ships – Already realized	Populations at military bases increase due to housing of quarantined populations; military facility is faced with unexpected enumerating duties	**Offer GQ Admins alternative methods of enumeration** • Exploring alternative methods of swearing in GQ admins for Self-Facility enumeration



OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Risks Being Monitored

#	Title	Description	Mitigation
7	Staffing not available to handle specific planned activities – Already realized	Expected mail out of letters and packages to GQ administrators to enable them to complete their enumeration process (eResponse). NPC working out logistics to be able to assist with responding to GQ Admin	**Offer alternative method of sending login credentials.** • MOJO HERMES Email Blasts using data captured and received from NPC ATAC ERDT • MOJO HERMES Email Blasts using information captured and received from FOCS in ACOs
8	Organizations serving SBE locations cease operations (e.g. mobile food vans) – Already realized	Organizations serving locations that target people experiencing homelessness are not allowed to operate	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
9	Hotels/Motels become quarantine facilities/ being used to house people experiencing homelessness – Already realized	Hotels/motels not previously considered TLs could become SBE locations or housing facility for quarantined people	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
10	TLs cease operations due to ban/quarantine (e.g. carnivals) – Already realized	City or State Government bans large crowds and gatherings due to exposure risk, such as parks, etc.	**Explore alternative methods of creating specific universe (e.g. carnival/circuses)** • Allow local knowledge to help with determining that universe. If there is a scheduled event, an appointment for enumeration would be set.



OD 8 GQE, SBE, ETL, MVE & Military Operational Outreach Throughout COVID

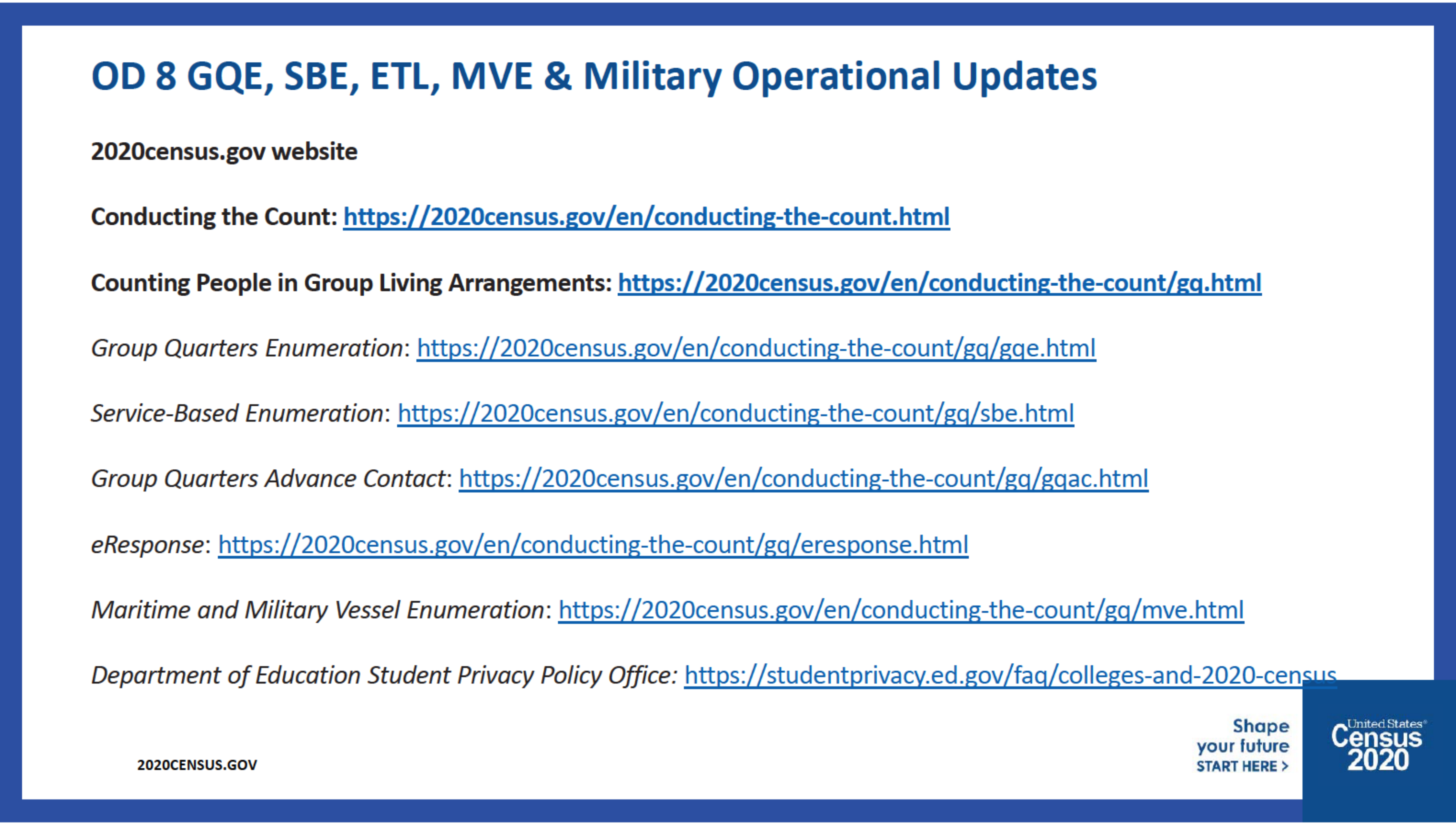
- March – Nationwide “Stay at Home” orders
- **March:** Letter developed by DCMD for RCCs to send to GQs that selected self enumeration options to change option to eResponse or paper listings
- **March 13:** Posted letter, *Update on 2020 Census for Student Housing Administrators* on the Census Bureau and Department of Education Website, requesting administrators who selected self enumeration option to change option to eResponse or Paper Listings
- **March 25:** DCMD sent *Update on the 2020 Census for Health Care Administrator* letter to Health Care umbrella organizations providing guidance for Administrators that selected a self-enumeration option to change method of enumeration to eResponse of Paper Listings
- NPC/ Jeffersonville call center was closed due to COVID and was not available to complete task in support of eResponse. As a result:
  - NPC was unable to meet the March 27 deadline for mailing eResponse Letters with Login credentials
  - DCMD worked with NPC ATAC management to update the system to allow multiple users to view and update email address
  - DCMD stood up a Call Center with staff across ACOs and Census HQ volunteers to verify/update email address to deliver login credentials.
- **March 31:** Mojo/Hermes sent out 1^{st} email blast with login credentials to GQ admins who has selected eResponse during GQAC
- **April 1:** GQE eResponse portal became available for GQ submittals.
- **April 13** and **April 20:** Mojo/Hermes sent out email blast 2 and 3 with login credentials for bounce back emails from 1^{st} email blast
- **April 2:** DCMD Staff and volunteers across the Decennial Directorate and other started reaching out in response to questions from GQ administrators received via email and phone calls.
- April ? PIO developed a video to college students. Posted on website encouraging internet response or that GQ admins would respond for them if they live in student housing
- **April 20:** ACO began calling GQ admins to offer Mail out/ Mail back Paper Response Data Collection (Paper Listings)
- **May 28:** Census Bureau participated in a webinar to remind/ update student housing administrators on the 2020 Census Group Quarters operation and to inform administrators of the upcoming request for off-campus student data.
- **June 3:** NPC mailed Maritime/Military Vessel reminder letters to non-responding vessel operators
- **June 8:** ACO staff began calling GQ administrators to reschedule appointment dates for their facilities.
- **June 11:** Meeting with AACRAO…
- **June 22:** Provided updates to The Salvation Army for their Directive to be sent to their managing entities.
- **June 22:** Met with National Network to End Domestic Violence to discuss upcoming enumeration, options, and COVID-19 procedures

21 Along 2020CENSUS.GOV

Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



OD 8 GQE, SBE, ETL, MVE & Military Operational Updates

2020census.gov website

Conducting the Count: https://2020census.gov/en/conducting-the-count.html

Counting People in Group Living Arrangements: https://2020census.gov/en/conducting-the-count/gq.html

Group Quarters Enumeration: https://2020census.gov/en/conducting-the-count/gq/gqe.html

Service-Based Enumeration: https://2020census.gov/en/conducting-the-count/gq/sbe.html

Group Quarters Advance Contact: https://2020census.gov/en/conducting-the-count/gq/gqac.html

eResponse: https://2020census.gov/en/conducting-the-count/gq/eresponse.html

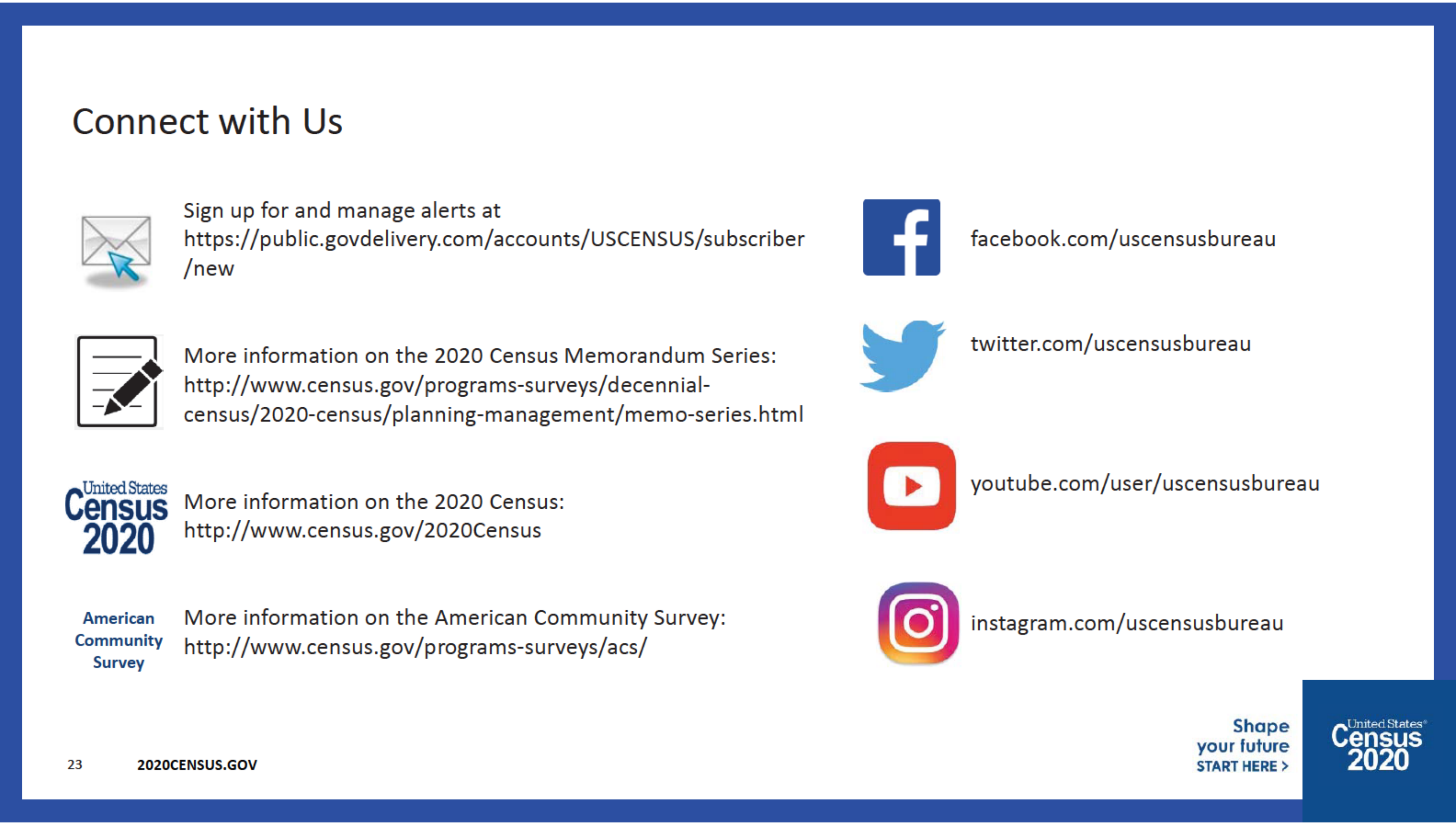
Maritime and Military Vessel Enumeration: https://2020census.gov/en/conducting-the-count/gq/mve.html

Department of Education Student Privacy Policy Office: https://studentprivacy.ed.gov/faq/colleges-and-2020-census

2020CENSUS.GOV

Shape
your future
START HERE >

United States Census 2020



Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

United States Census 2020

More information on the 2020 Census: http://www.census.gov/2020Census

American Community Survey

More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

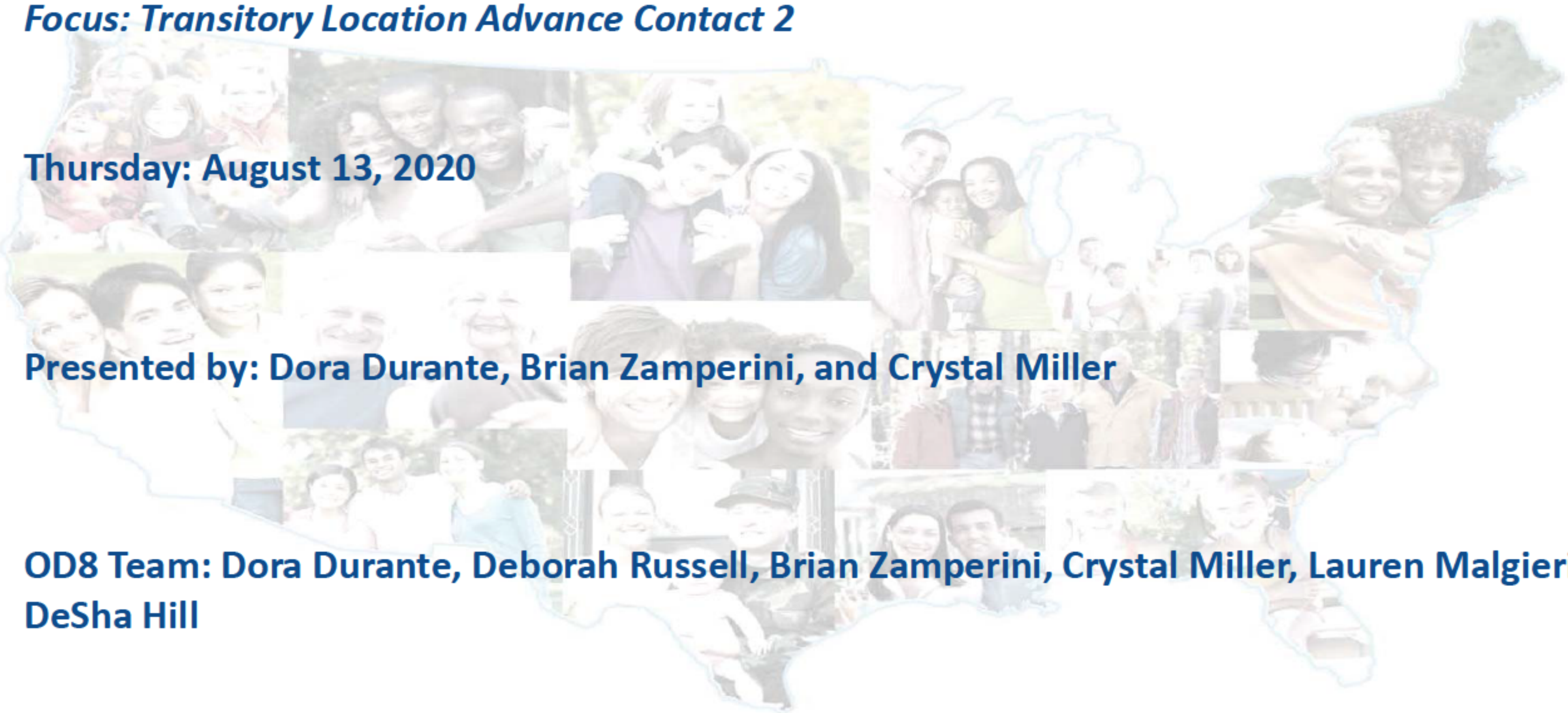
instagram.com/uscensusbureau

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Operational Delivery 8: Transitory Locations



Focus: Transitory Location Advance Contact 2

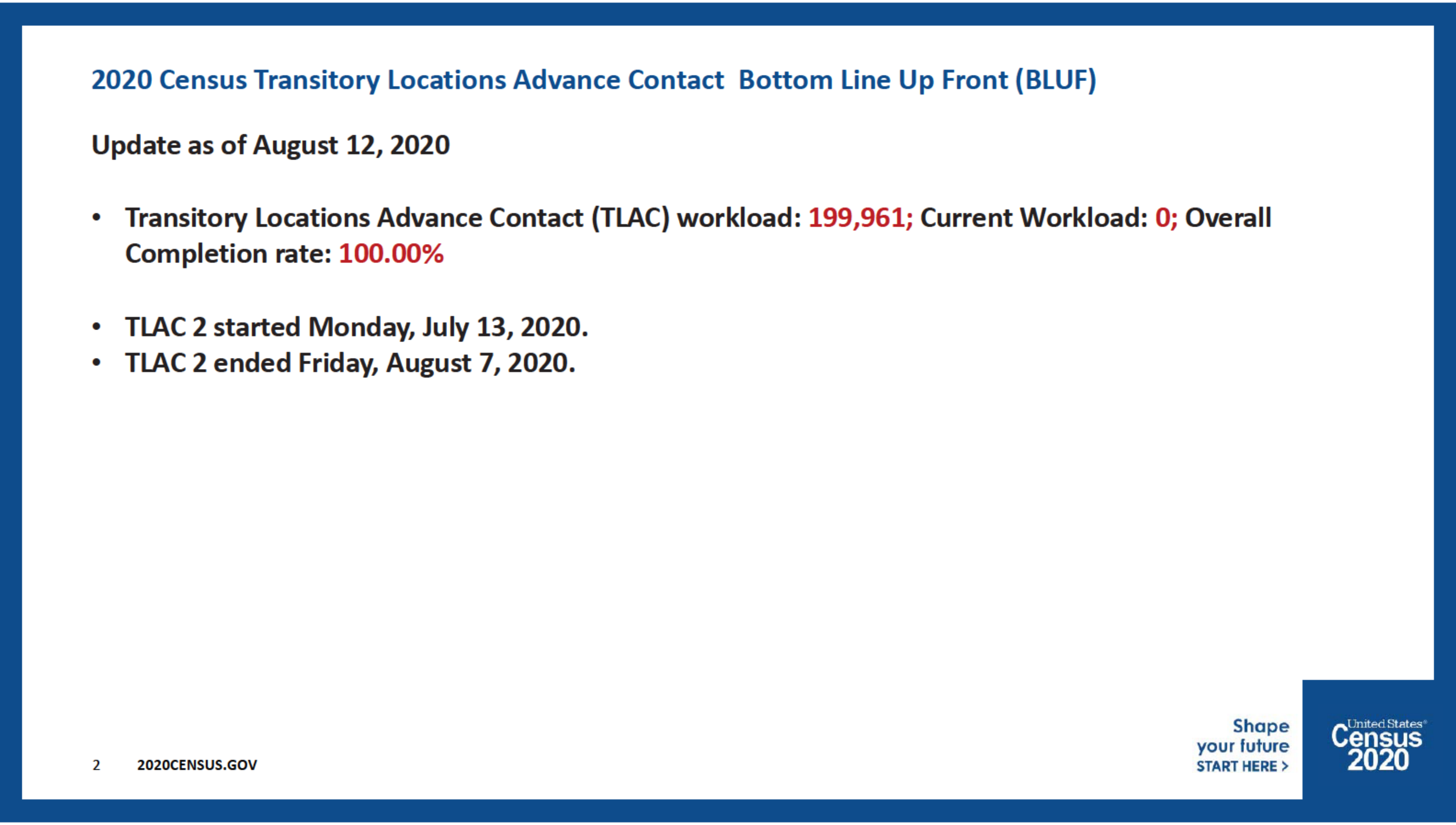
Thursday: August 13, 2020

Presented by: Dora Durante, Brian Zamperini, and Crystal Miller

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

Shape
your future
START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census Transitory Locations Advance Contact Bottom Line Up Front (BLUF)

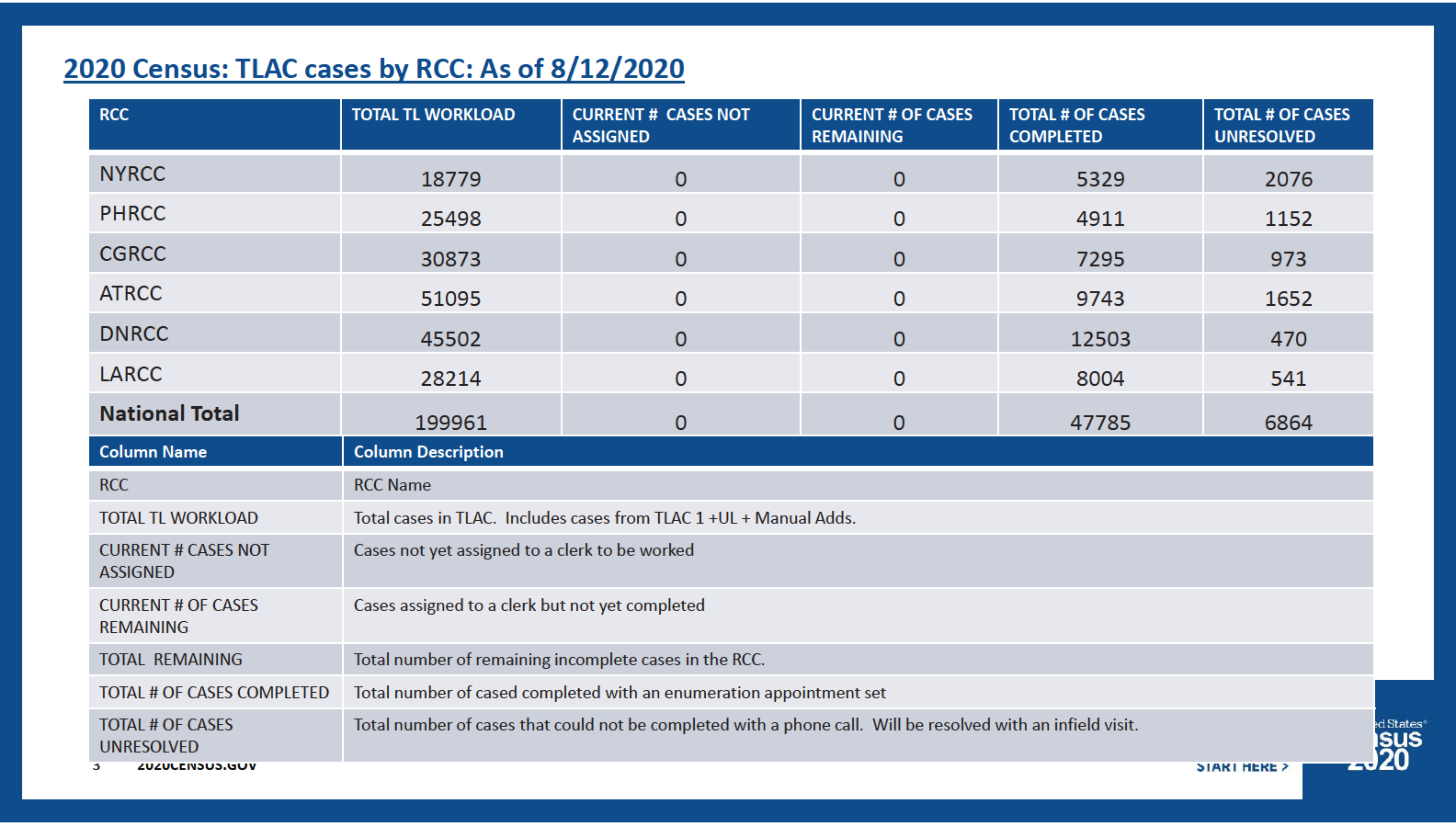
Update as of August 12, 2020

- **Transitory Locations Advance Contact (TLAC) workload: 199,961; Current Workload: 0; Overall Completion rate: 100.00%**

- **TLAC 2 started Monday, July 13, 2020.**
- **TLAC 2 ended Friday, August 7, 2020.**

Shape
your future
START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census: TLAC cases by RCC: As of 8/12/2020

RCC	TOTAL TL WORKLOAD	CURRENT # CASES NOT ASSIGNED	CURRENT # OF CASES REMAINING	TOTAL # OF CASES COMPLETED	TOTAL # OF CASES UNRESOLVED
NYRCC	18779	0	0	5329	2076
PHRCC	25498	0	0	4911	1152
CGRCC	30873	0	0	7295	973
ATRCC	51095	0	0	9743	1652
DNRCC	45502	0	0	12503	470
LARCC	28214	0	0	8004	541
National Total	199961	0	0	47785	6864

Column Name	Column Description
RCC	RCC Name
TOTAL TL WORKLOAD	Total cases in TLAC. Includes cases from TLAC 1 +UL + Manual Adds.
CURRENT # CASES NOT ASSIGNED	Cases not yet assigned to a clerk to be worked
CURRENT # OF CASES REMAINING	Cases assigned to a clerk but not yet completed
TOTAL REMAINING	Total number of remaining incomplete cases in the RCC.
TOTAL # OF CASES COMPLETED	Total number of cased completed with an enumeration appointment set
TOTAL # OF CASES UNRESOLVED	Total number of cases that could not be completed with a phone call. Will be resolved with an infield visit.

3 2020CENSUS.GOV

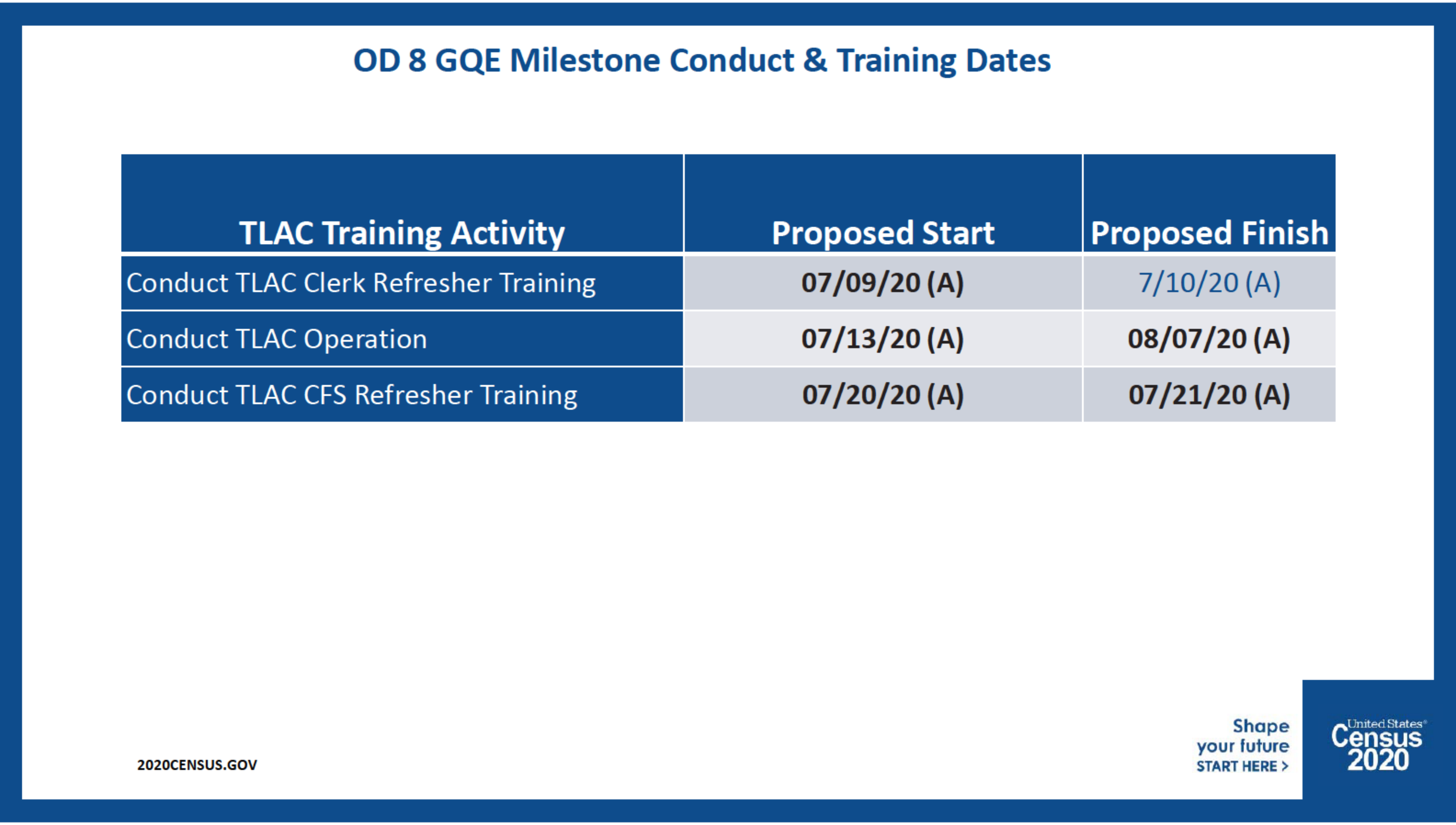
START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002

Backup

Schedule, Staffing, Risks

DRB Approval Number: CBDRB-FY21-DSEP-002



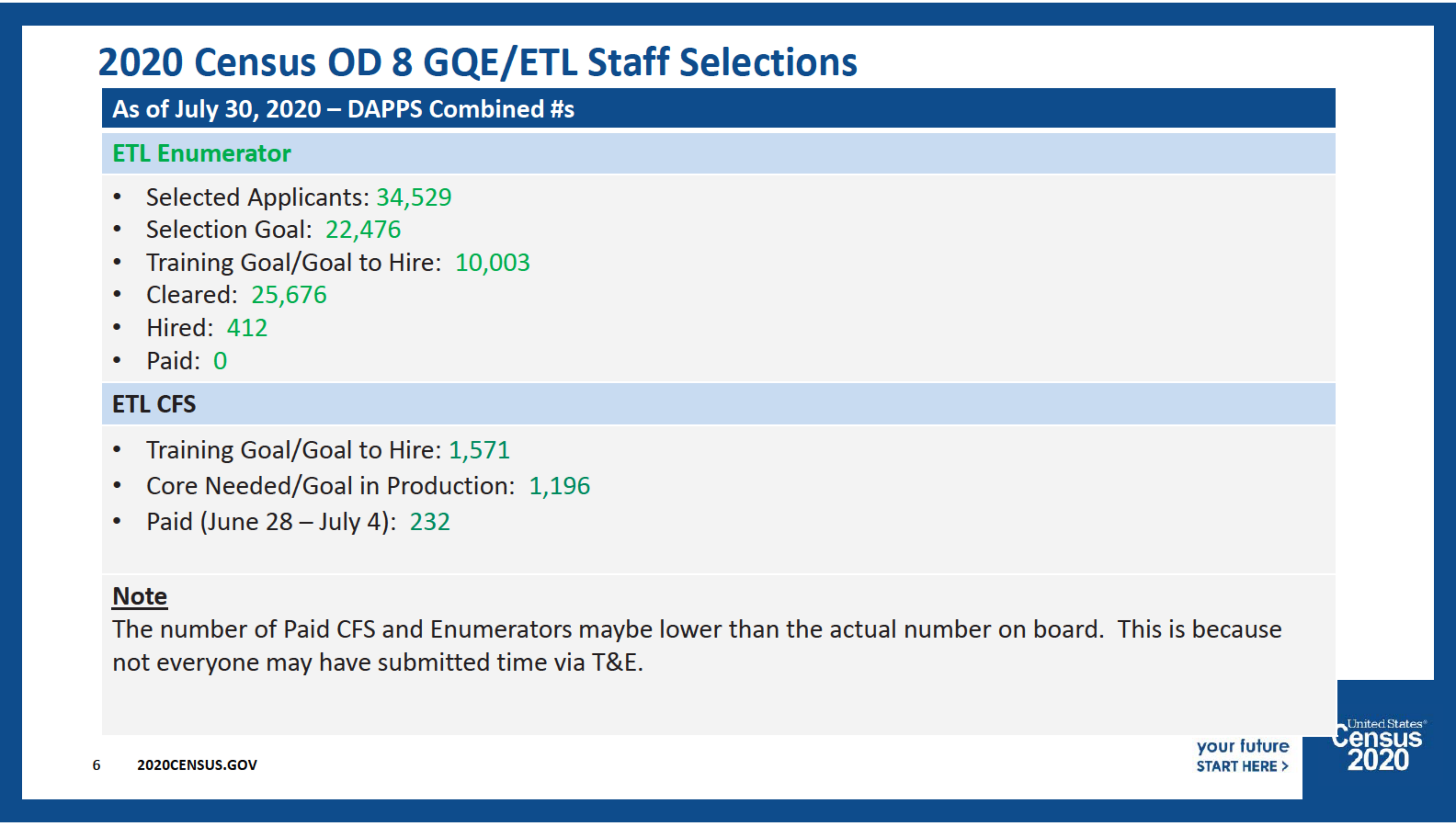
OD 8 GQE Milestone Conduct & Training Dates

TLAC Training Activity	Proposed Start	Proposed Finish
Conduct TLAC Clerk Refresher Training	07/09/20 (A)	7/10/20 (A)
Conduct TLAC Operation	07/13/20 (A)	08/07/20 (A)
Conduct TLAC CFS Refresher Training	07/20/20 (A)	07/21/20 (A)

2020CENSUS.GOV

Shape
your future
START HERE >

United States
Census
2020



2020 Census OD 8 GQE/ETL Staff Selections

As of July 30, 2020 – DAPPS Combined #s

ETL Enumerator

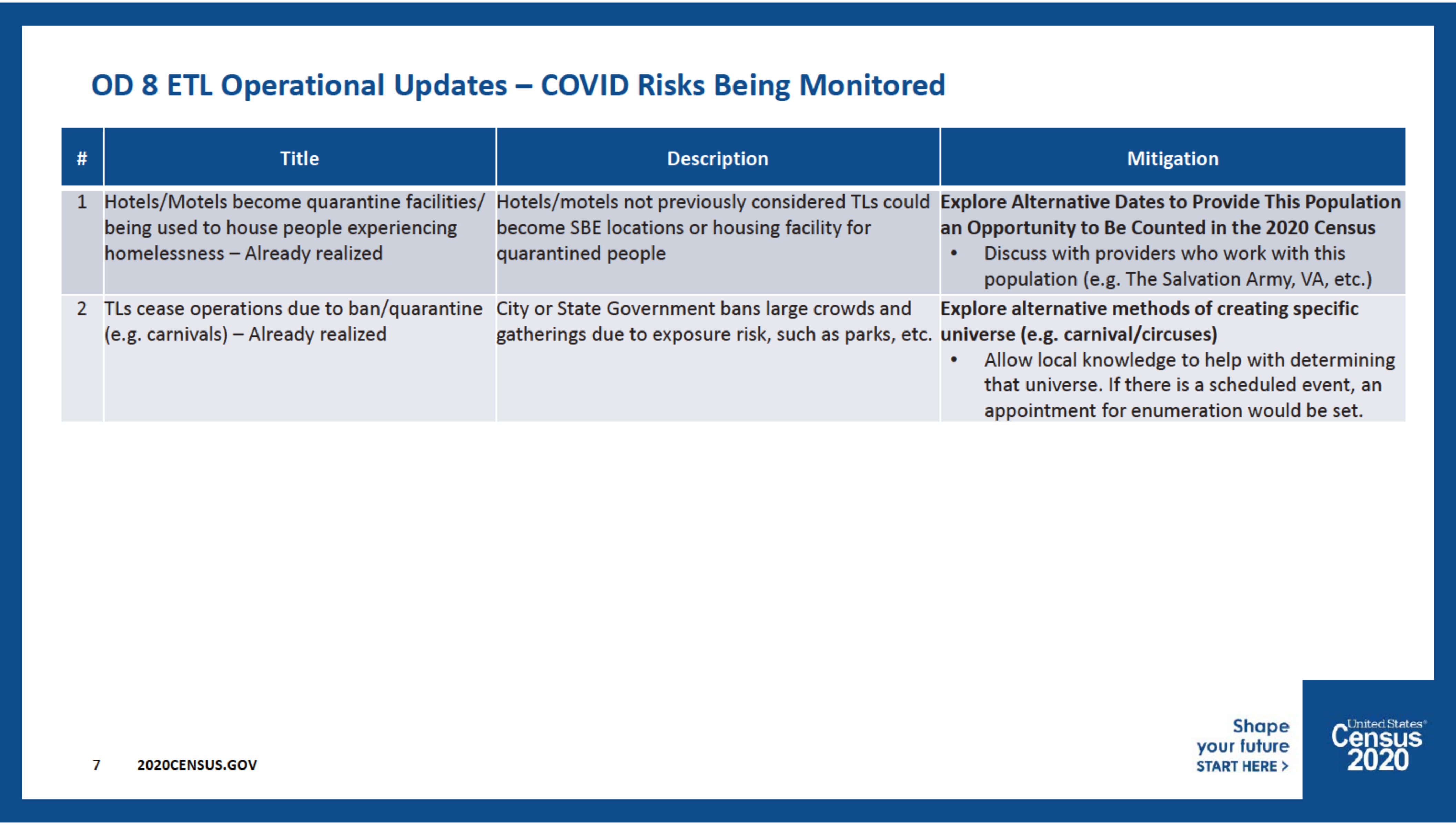
- Selected Applicants: 34,529
- Selection Goal: 22,476
- Training Goal/Goal to Hire: 10,003
- Cleared: 25,676
- Hired: 412
- Paid: 0

ETL CFS

- Training Goal/Goal to Hire: 1,571
- Core Needed/Goal in Production: 1,196
- Paid (June 28 – July 4): 232

Note

The number of Paid CFS and Enumerators maybe lower than the actual number on board. This is because not everyone may have submitted time via T&E.

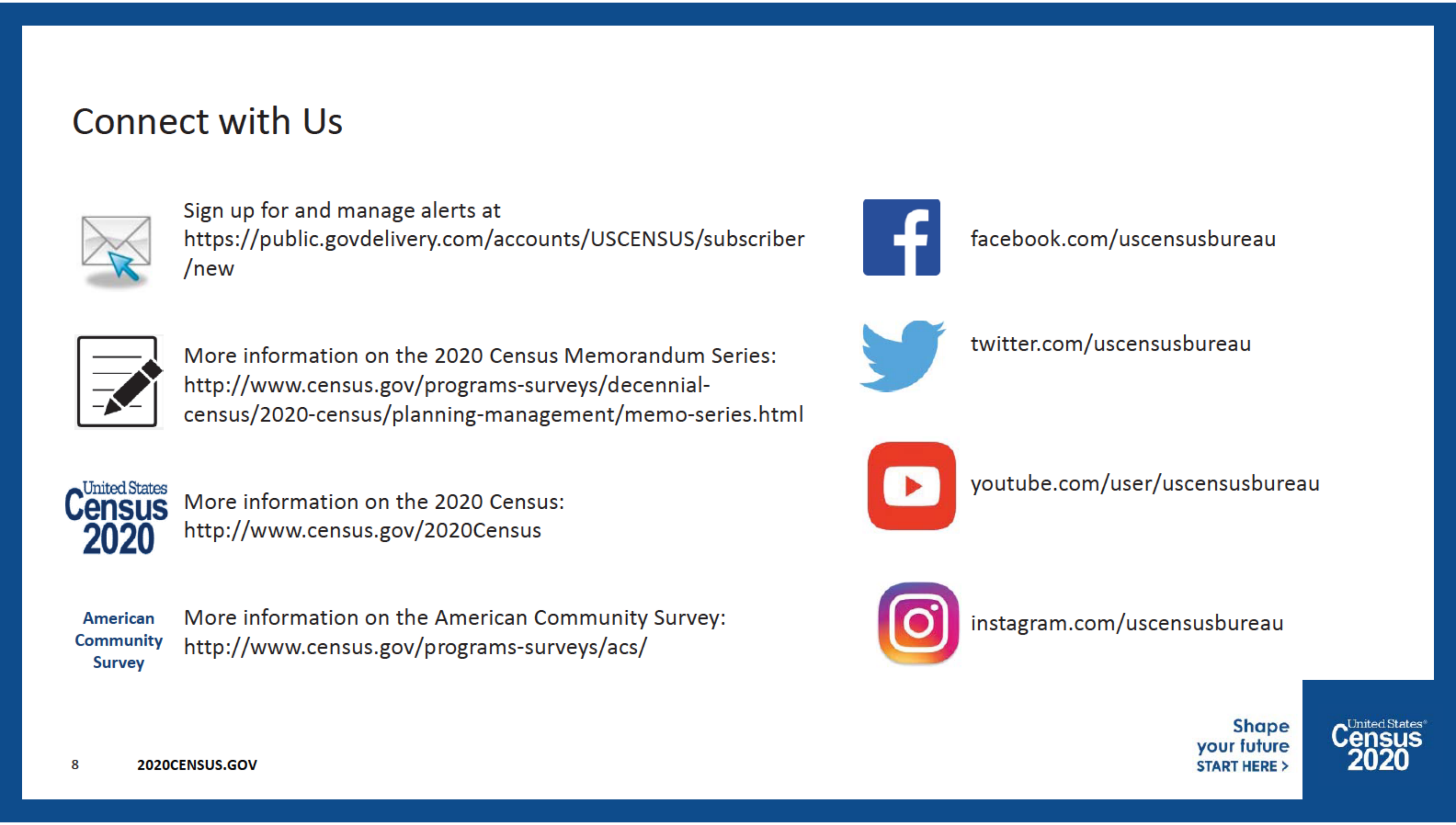


OD 8 ETL Operational Updates – COVID Risks Being Monitored

#	Title	Description	Mitigation
1	Hotels/Motels become quarantine facilities/ being used to house people experiencing homelessness – Already realized	Hotels/motels not previously considered TLs could become SBE locations or housing facility for quarantined people	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
2	TLs cease operations due to ban/quarantine (e.g. carnivals) – Already realized	City or State Government bans large crowds and gatherings due to exposure risk, such as parks, etc.	**Explore alternative methods of creating specific universe (e.g. carnival/circuses)** • Allow local knowledge to help with determining that universe. If there is a scheduled event, an appointment for enumeration would be set.

Shape
your future
START HERE >

United States®
Census
2020



Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

United States Census 2020 More information on the 2020 Census: http://www.census.gov/2020Census

American Community Survey More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

instagram.com/uscensusbureau

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Operational Delivery 8: Group Quarters Update

Focus: Group Quarters Enumeration and Maritime/Military Vessel Enumeration

Thursday: August 27, 2020

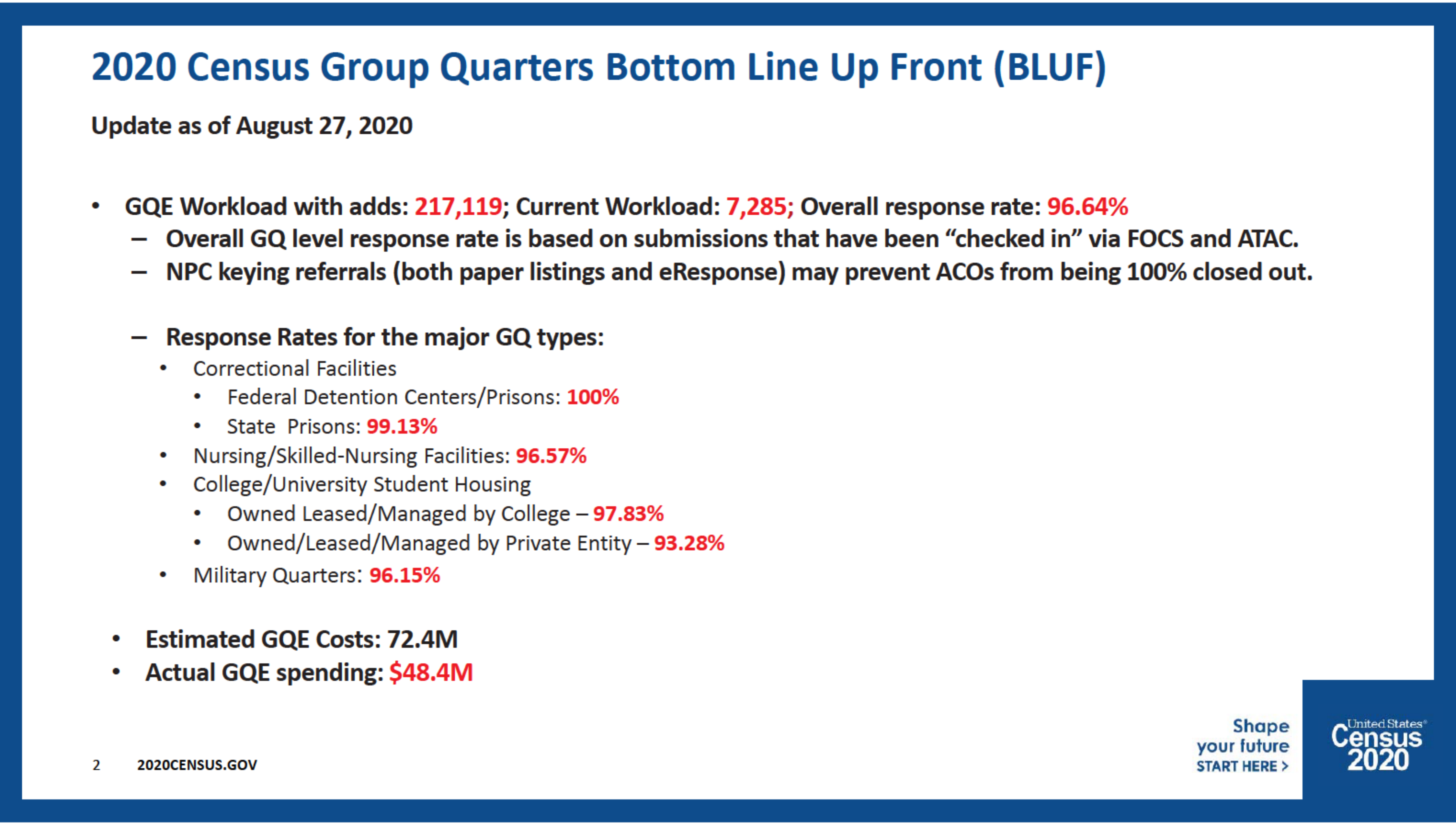
Presented by: Judy Belton, Deborah Russell and Crystal Miller

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

Shape
your future
START HERE >

United States®
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002



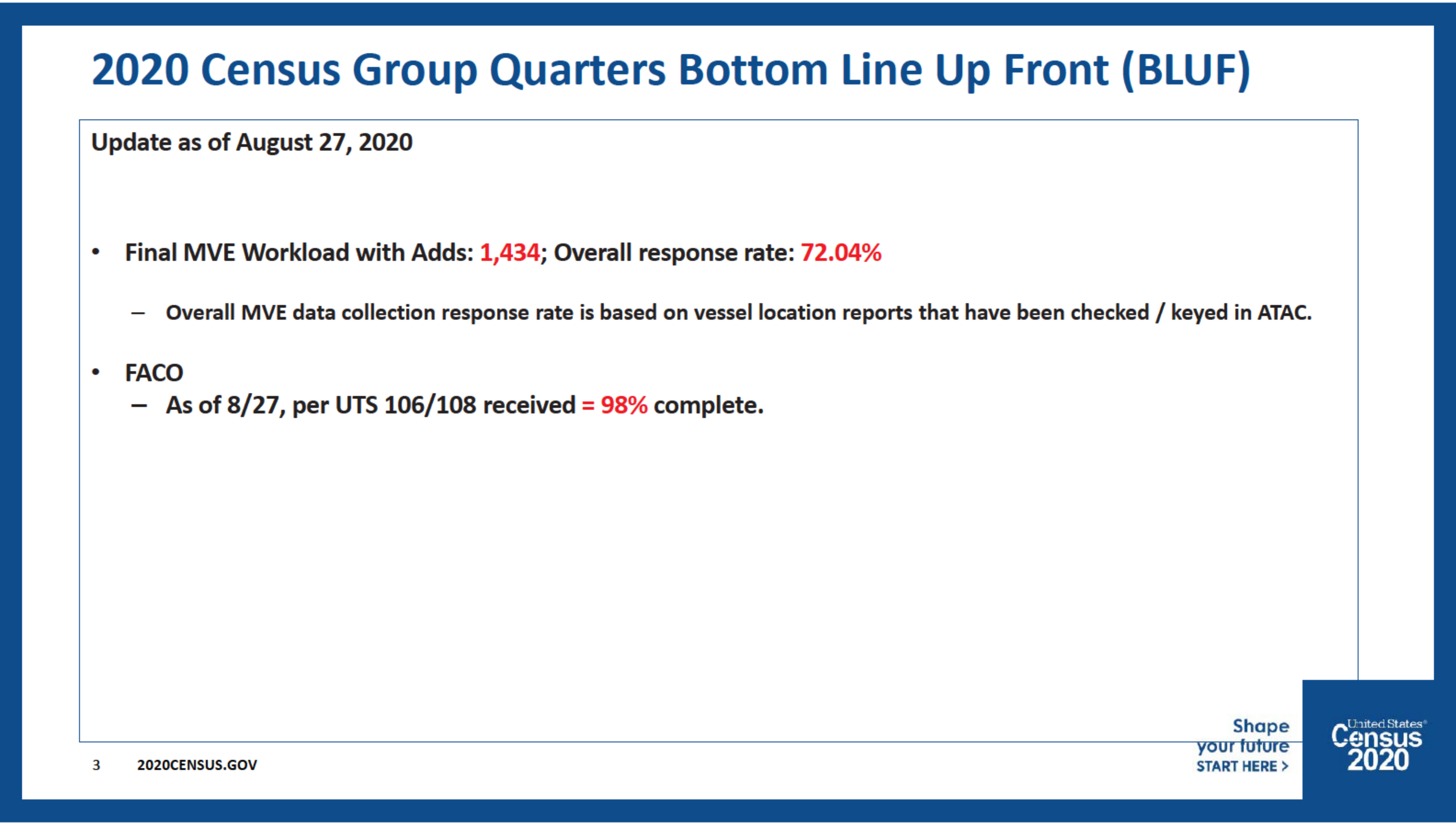
2020 Census Group Quarters Bottom Line Up Front (BLUF)

Update as of August 27, 2020

- **GQE Workload with adds: 217,119; Current Workload: 7,285; Overall response rate: 96.64%**
  - **Overall GQ level response rate is based on submissions that have been "checked in" via FOCS and ATAC.**
  - **NPC keying referrals (both paper listings and eResponse) may prevent ACOs from being 100% closed out.**
  - **Response Rates for the major GQ types:**
    - Correctional Facilities
      - Federal Detention Centers/Prisons: **100%**
      - State Prisons: **99.13%**
    - Nursing/Skilled-Nursing Facilities: **96.57%**
    - College/University Student Housing
      - Owned Leased/Managed by College – **97.83%**
      - Owned/Leased/Managed by Private Entity – **93.28%**
    - Military Quarters: **96.15%**
- **Estimated GQE Costs: 72.4M**
- **Actual GQE spending: $48.4M**

Shape
your future
START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



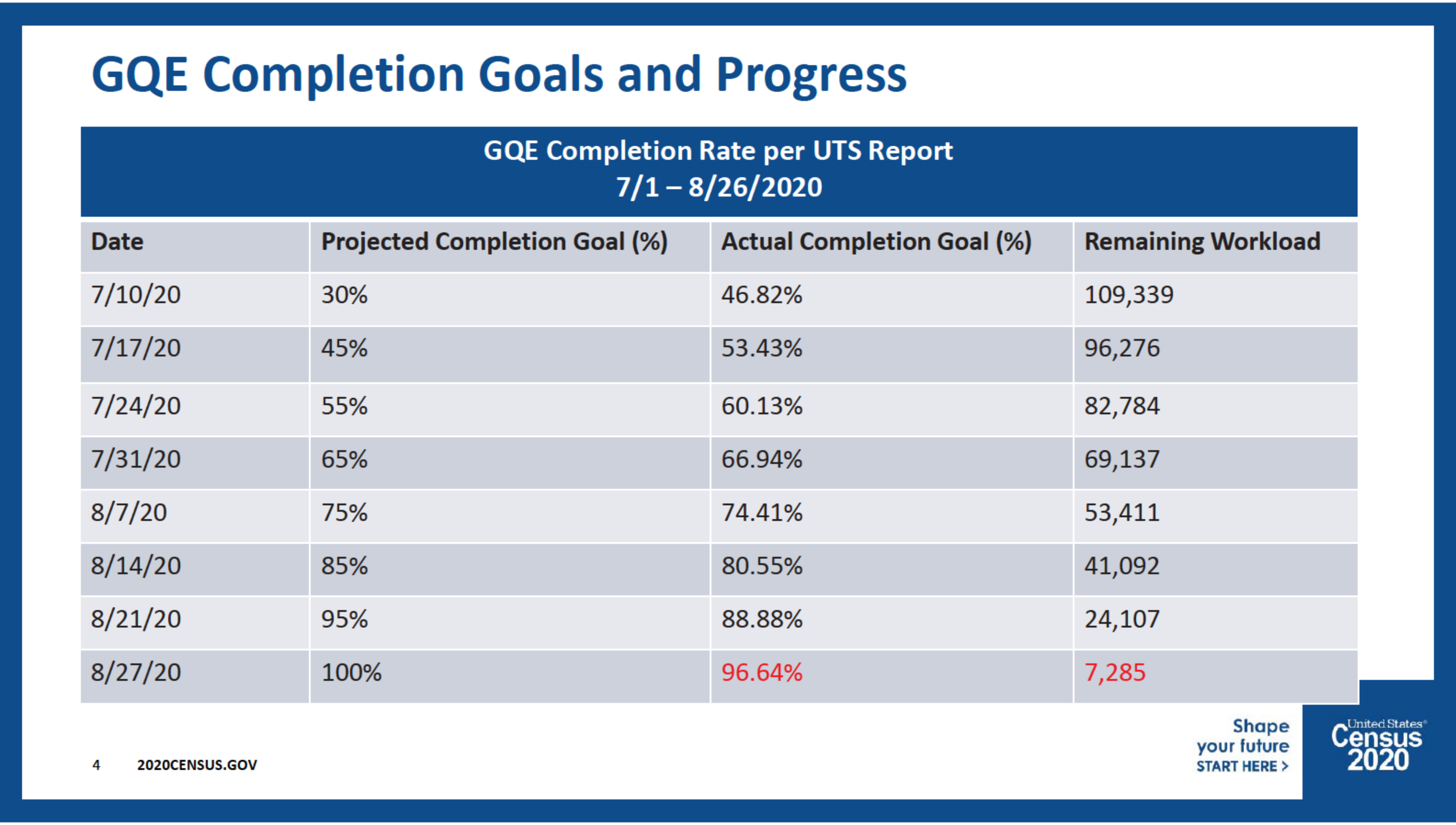
2020 Census Group Quarters Bottom Line Up Front (BLUF)

Update as of August 27, 2020

- **Final MVE Workload with Adds: 1,434; Overall response rate: 72.04%**
  - **Overall MVE data collection response rate is based on vessel location reports that have been checked / keyed in ATAC.**
- **FACO**
  - **As of 8/27, per UTS 106/108 received = 98% complete.**

Shape your future START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



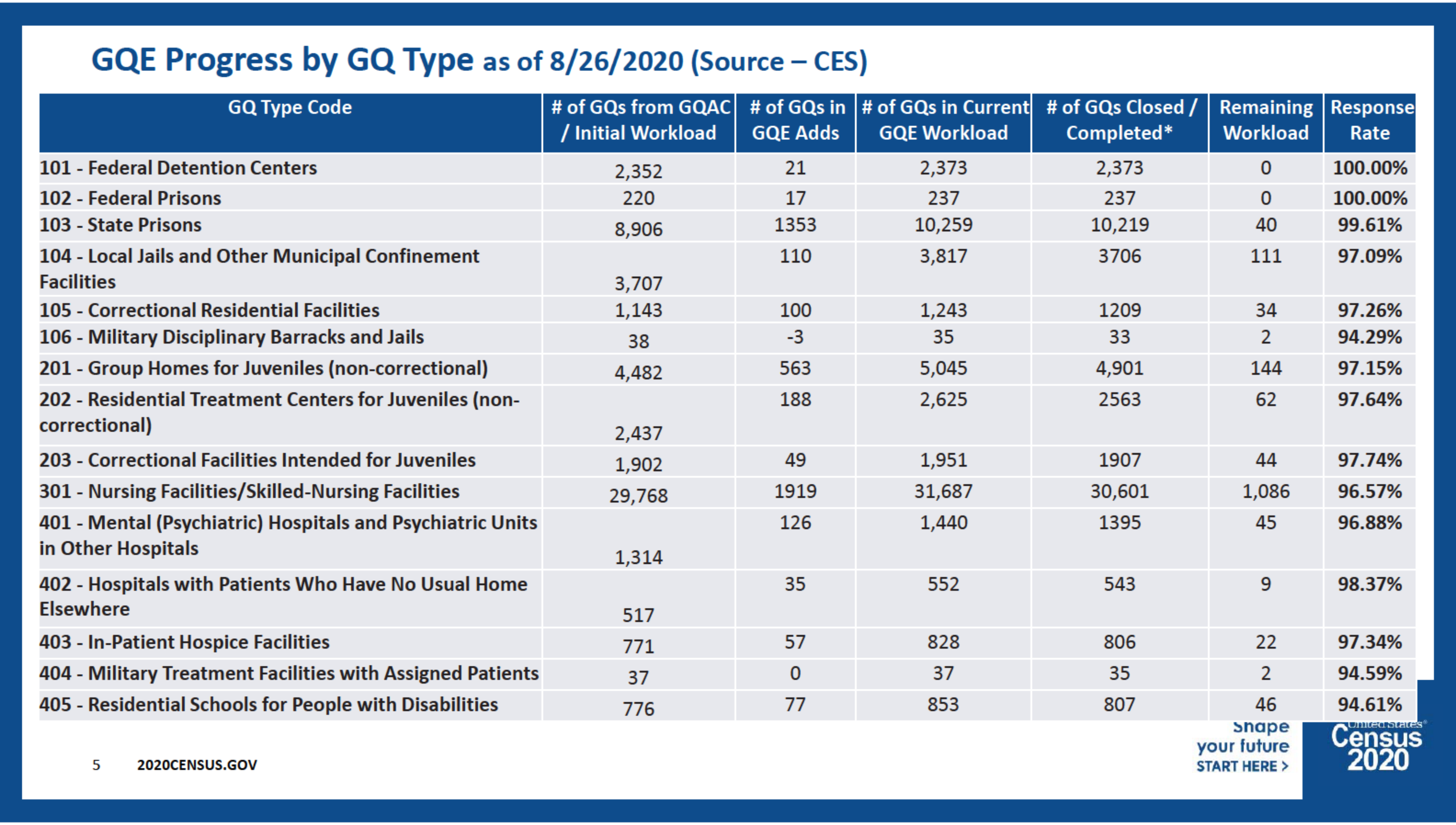
GQE Completion Goals and Progress

GQE Completion Rate per UTS Report 7/1 – 8/26/2020			
Date	**Projected Completion Goal (%)**	**Actual Completion Goal (%)**	**Remaining Workload**
7/10/20	30%	46.82%	109,339
7/17/20	45%	53.43%	96,276
7/24/20	55%	60.13%	82,784
7/31/20	65%	66.94%	69,137
8/7/20	75%	74.41%	53,411
8/14/20	85%	80.55%	41,092
8/21/20	95%	88.88%	24,107
8/27/20	100%	96.64%	7,285

Shape
your future
START HERE >

United States
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002



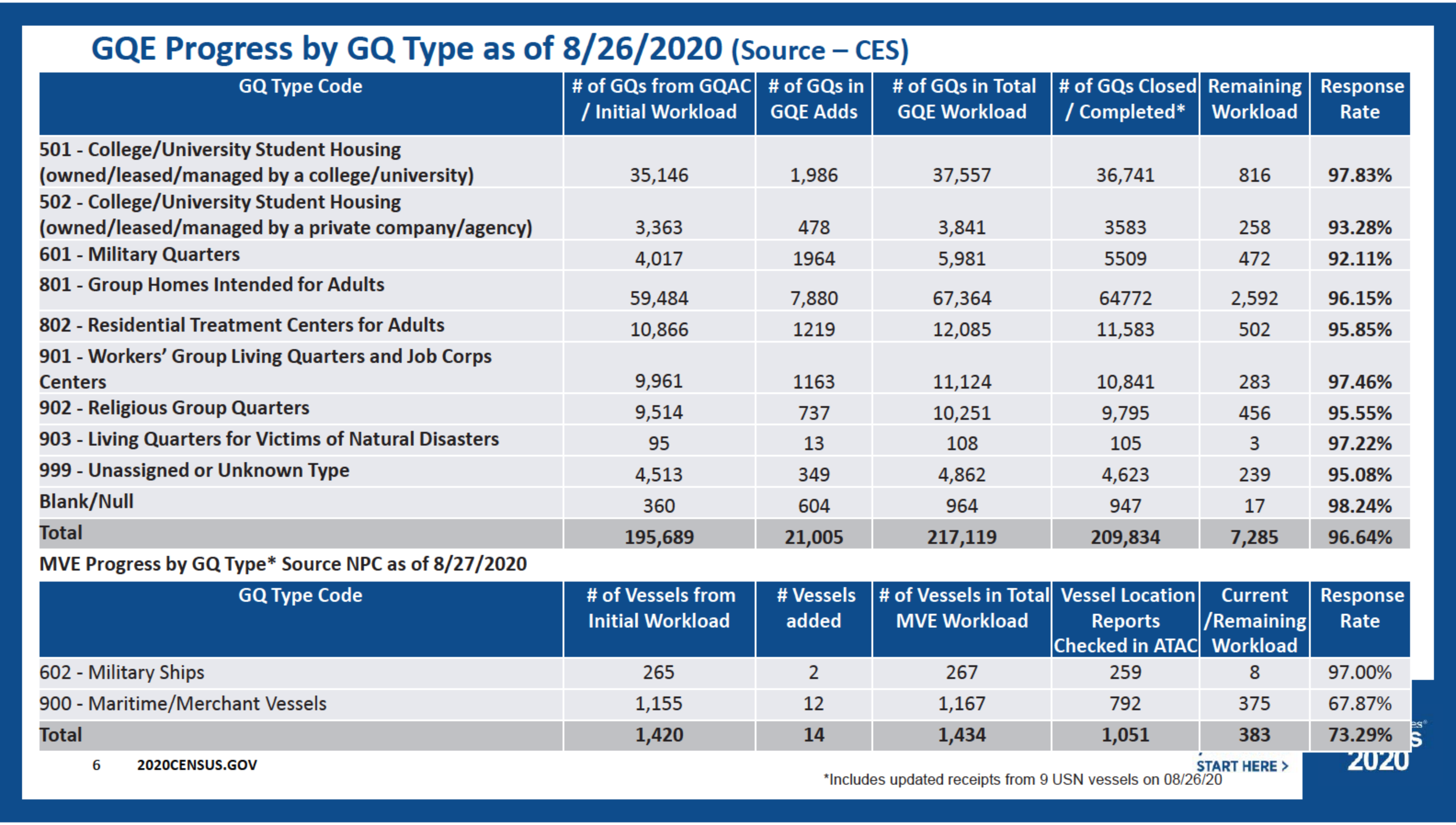
GQE Progress by GQ Type as of 8/26/2020 (Source – CES)

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Current GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
101 - Federal Detention Centers	2,352	21	2,373	2,373	0	100.00%
102 - Federal Prisons	220	17	237	237	0	100.00%
103 - State Prisons	8,906	1353	10,259	10,219	40	99.61%
104 - Local Jails and Other Municipal Confinement Facilities	3,707	110	3,817	3706	111	97.09%
105 - Correctional Residential Facilities	1,143	100	1,243	1209	34	97.26%
106 - Military Disciplinary Barracks and Jails	38	-3	35	33	2	94.29%
201 - Group Homes for Juveniles (non-correctional)	4,482	563	5,045	4,901	144	97.15%
202 - Residential Treatment Centers for Juveniles (non-correctional)	2,437	188	2,625	2563	62	97.64%
203 - Correctional Facilities Intended for Juveniles	1,902	49	1,951	1907	44	97.74%
301 - Nursing Facilities/Skilled-Nursing Facilities	29,768	1919	31,687	30,601	1,086	96.57%
401 - Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals	1,314	126	1,440	1395	45	96.88%
402 - Hospitals with Patients Who Have No Usual Home Elsewhere	517	35	552	543	9	98.37%
403 - In-Patient Hospice Facilities	771	57	828	806	22	97.34%
404 - Military Treatment Facilities with Assigned Patients	37	0	37	35	2	94.59%
405 - Residential Schools for People with Disabilities	776	77	853	807	46	94.61%

Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



GQE Progress by GQ Type as of 8/26/2020 (Source – CES)

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Total GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
501 - College/University Student Housing (owned/leased/managed by a college/university)	35,146	1,986	37,557	36,741	816	97.83%
502 - College/University Student Housing (owned/leased/managed by a private company/agency)	3,363	478	3,841	3583	258	93.28%
601 - Military Quarters	4,017	1964	5,981	5509	472	92.11%
801 - Group Homes Intended for Adults	59,484	7,880	67,364	64772	2,592	96.15%
802 - Residential Treatment Centers for Adults	10,866	1219	12,085	11,583	502	95.85%
901 - Workers' Group Living Quarters and Job Corps Centers	9,961	1163	11,124	10,841	283	97.46%
902 - Religious Group Quarters	9,514	737	10,251	9,795	456	95.55%
903 - Living Quarters for Victims of Natural Disasters	95	13	108	105	3	97.22%
999 - Unassigned or Unknown Type	4,513	349	4,862	4,623	239	95.08%
Blank/Null	360	604	964	947	17	98.24%
Total	195,689	21,005	217,119	209,834	7,285	96.64%

MVE Progress by GQ Type* Source NPC as of 8/27/2020

GQ Type Code	# of Vessels from Initial Workload	# Vessels added	# of Vessels in Total MVE Workload	Vessel Location Reports Checked in ATAC	Current /Remaining Workload	Response Rate
602 - Military Ships	265	2	267	259	8	97.00%
900 - Maritime/Merchant Vessels	1,155	12	1,167	792	375	67.87%
Total	1,420	14	1,434	1,051	383	73.29%

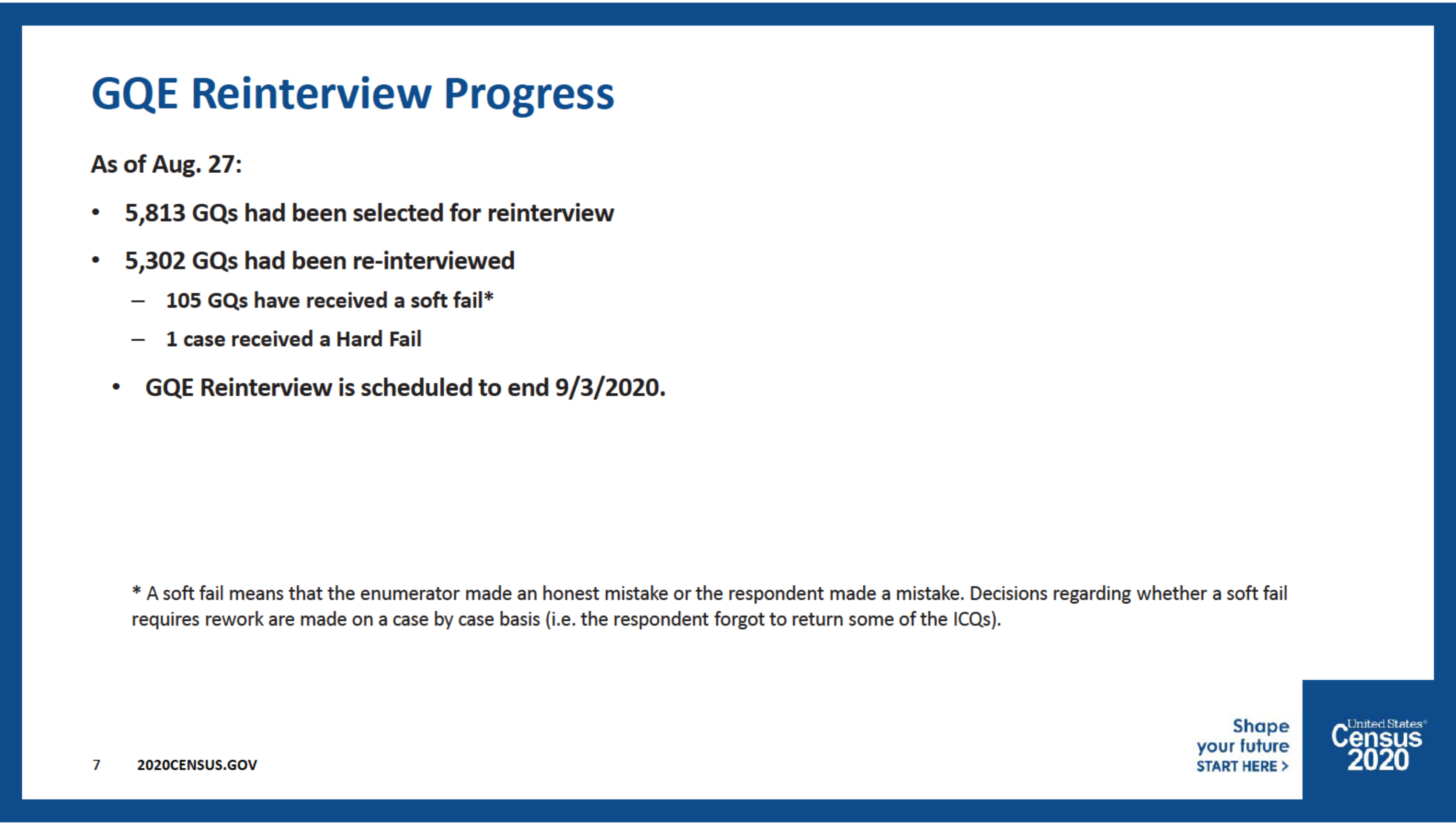
6 2020CENSUS.GOV

START HERE >

2020

*Includes updated receipts from 9 USN vessels on 08/26/20

DRB Approval Number: CBDRB-FY21-DSEP-002



GQE Reinterview Progress

As of Aug. 27:

- **5,813 GQs had been selected for reinterview**
- **5,302 GQs had been re-interviewed**
  - **105 GQs have received a soft fail***
  - **1 case received a Hard Fail**
- **GQE Reinterview is scheduled to end 9/3/2020.**

* A soft fail means that the enumerator made an honest mistake or the respondent made a mistake. Decisions regarding whether a soft fail requires rework are made on a case by case basis (i.e. the respondent forgot to return some of the ICQs).

Shape
your future
START HERE >

United States
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002

Backup

Schedules and Systems

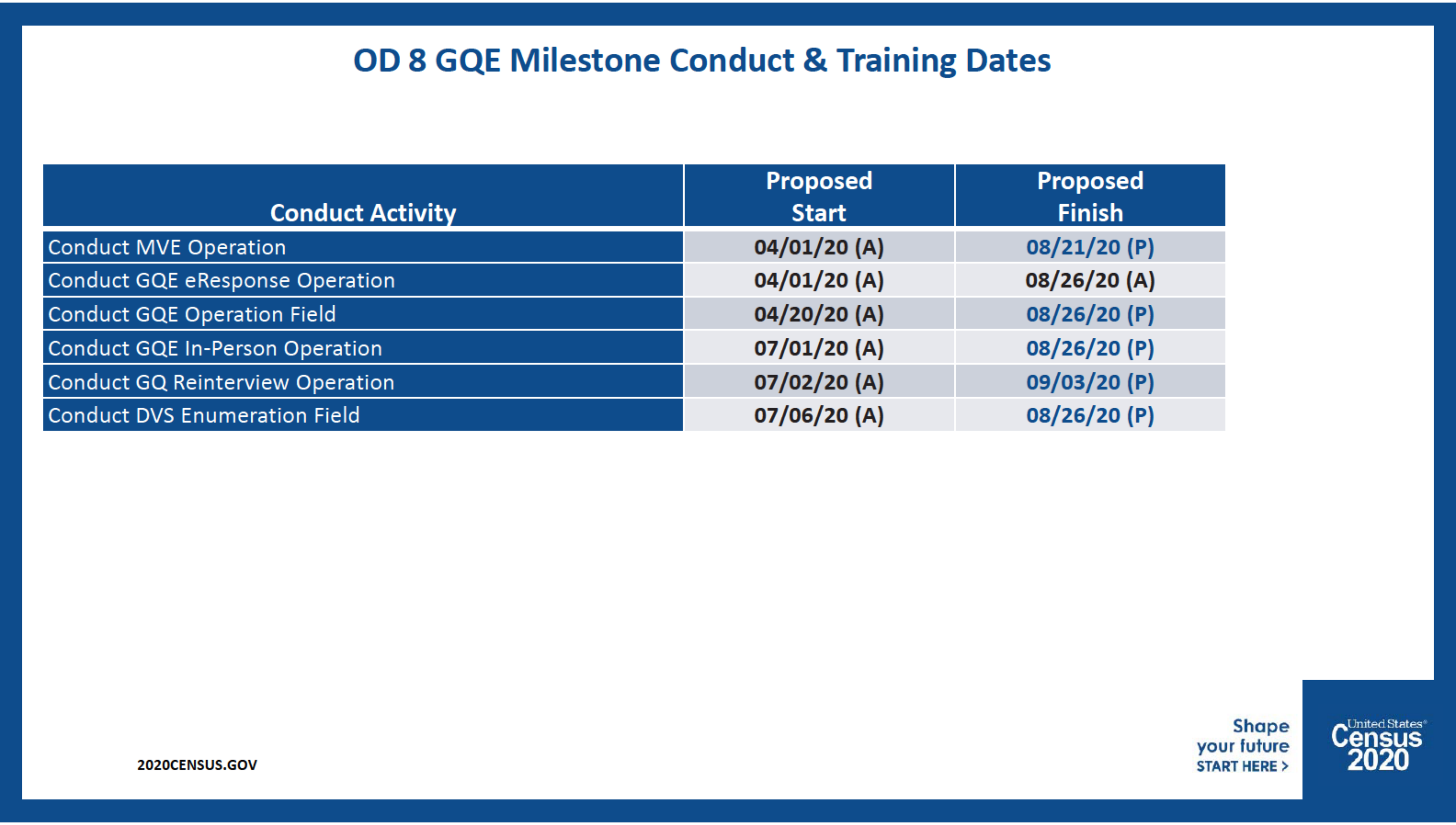
DRB Approval Number: CBDRB-FY21-DSEP-002

Maritime/Military Vessel Enumeration Progress

Maritime/ Military Agency	MVE Vessels in Agency	Checked In	% Vessels Checked into ATAC	# Checked-In Vessels Out of Scope	% Checked-In Vessels Out of Scope	# Vessels Enumerated	% Vessels Enumerated	# of In Progress Vessels	% of In Progress Vessels
National	1,434	1,033	72.04%	426	29.71%	419	29.22%	1,015	70.78%
CFEC	645	439	68.06%	289	44.81%	251	38.91%	394	61.09%
GRNC	5	5	100.00%	1	20.00%	1	20.00%	4	80.00%
LCA	50	46	92.00%	20	40.00%	19	38.00%	31	62.00%
MARAD	235	141	60.00%	32	13.62%	13	5.53%	222	94.47%
MSC	111	71	63.96%	23	20.72%	8	7.21%	103	92.79%
NMFS	74	44	59.46%	21	28.38%	17	22.97%	57	77.03%
NOAA	15	15	100.00%	7	46.67%	8	53.33%	7	46.67%
Other Maritime/ Military	16	14	87.50%	6	37.50%	4	25.00%	12	75.00%
UNOLS	18	17	94.44%	6	33.33%	5	27.78%	13	72.22%
USCG	59	46	77.97%	4	6.78%	25	42.37%	34	57.63%
USN	206	203	98.54%	17	8.25%	68	33.01%	138	66.99%

Source: UTS Report 8/26/2020 *includes 9 additional USN Vessels who submitted 08/26/20
*Per UTS, Total MVQs is 79,126

DRB Approval Number: CBDRB-FY21-DSEP-002



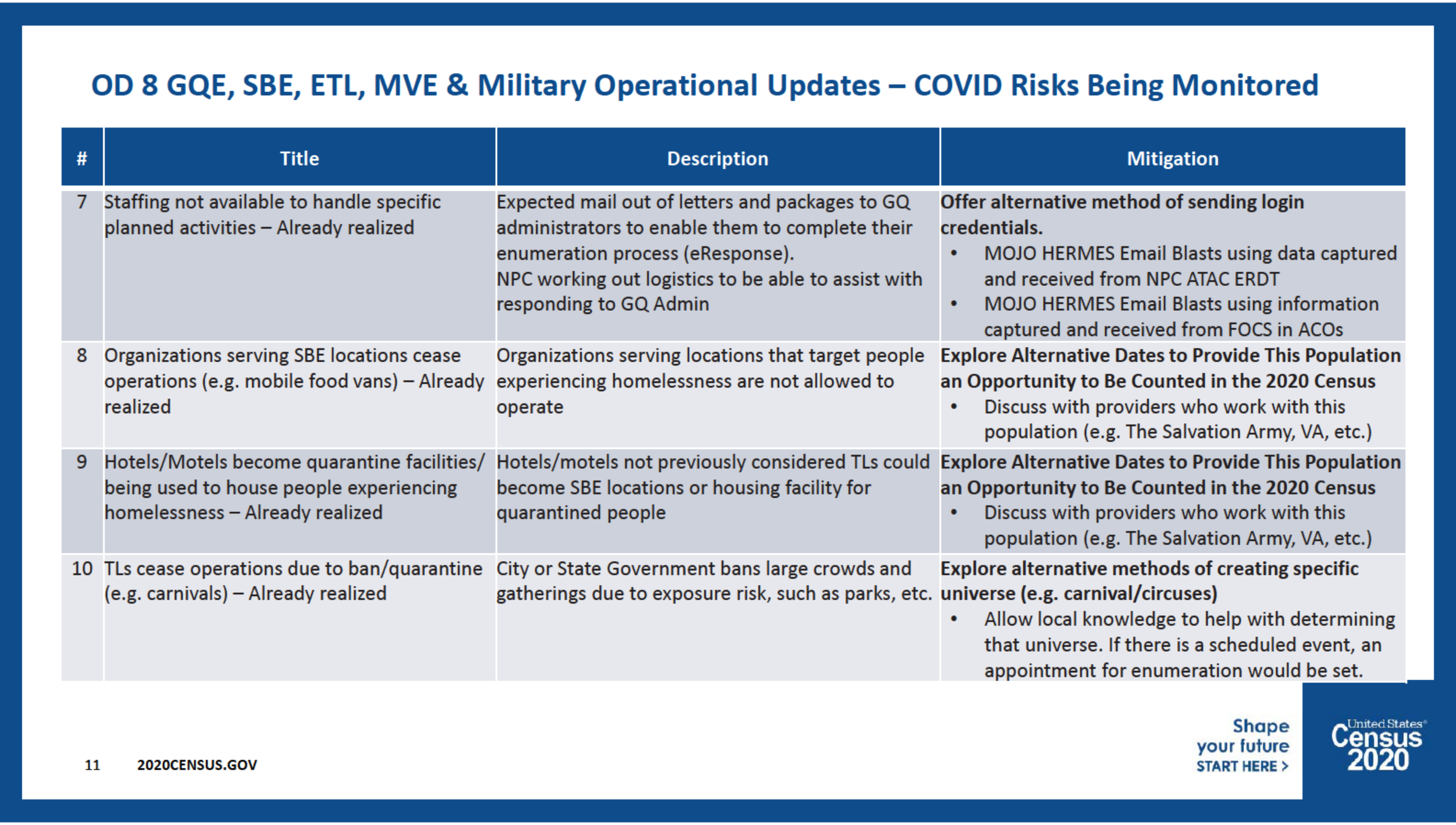
OD 8 GQE Milestone Conduct & Training Dates

Conduct Activity	Proposed Start	Proposed Finish
Conduct MVE Operation	04/01/20 (A)	08/21/20 (P)
Conduct GQE eResponse Operation	04/01/20 (A)	08/26/20 (A)
Conduct GQE Operation Field	04/20/20 (A)	08/26/20 (P)
Conduct GQE In-Person Operation	07/01/20 (A)	08/26/20 (P)
Conduct GQ Reinterview Operation	07/02/20 (A)	09/03/20 (P)
Conduct DVS Enumeration Field	07/06/20 (A)	08/26/20 (P)

2020CENSUS.GOV

Shape
your future
START HERE >

DRB Approval Number: CBDRB-FY21-DSEP-002



OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Risks Being Monitored

#	Title	Description	Mitigation
7	Staffing not available to handle specific planned activities – Already realized	Expected mail out of letters and packages to GQ administrators to enable them to complete their enumeration process (eResponse). NPC working out logistics to be able to assist with responding to GQ Admin	**Offer alternative method of sending login credentials.** • MOJO HERMES Email Blasts using data captured and received from NPC ATAC ERDT • MOJO HERMES Email Blasts using information captured and received from FOCS in ACOs
8	Organizations serving SBE locations cease operations (e.g. mobile food vans) – Already realized	Organizations serving locations that target people experiencing homelessness are not allowed to operate	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
9	Hotels/Motels become quarantine facilities/ being used to house people experiencing homelessness – Already realized	Hotels/motels not previously considered TLs could become SBE locations or housing facility for quarantined people	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
10	TLs cease operations due to ban/quarantine (e.g. carnivals) – Already realized	City or State Government bans large crowds and gatherings due to exposure risk, such as parks, etc.	**Explore alternative methods of creating specific universe (e.g. carnival/circuses)** • Allow local knowledge to help with determining that universe. If there is a scheduled event, an appointment for enumeration would be set.

DRB Approval Number: CBDRB-FY21-DSEP-002

OD 8 GQE, MVE & Military Operational Updates – 2020 Census Websites

2020census.gov website

Conducting the Count: https://2020census.gov/en/conducting-the-count.html

Counting People in Group Living Arrangements: https://2020census.gov/en/conducting-the-count/gq.html

Group Quarters Enumeration: https://2020census.gov/en/conducting-the-count/gq/gqe.html

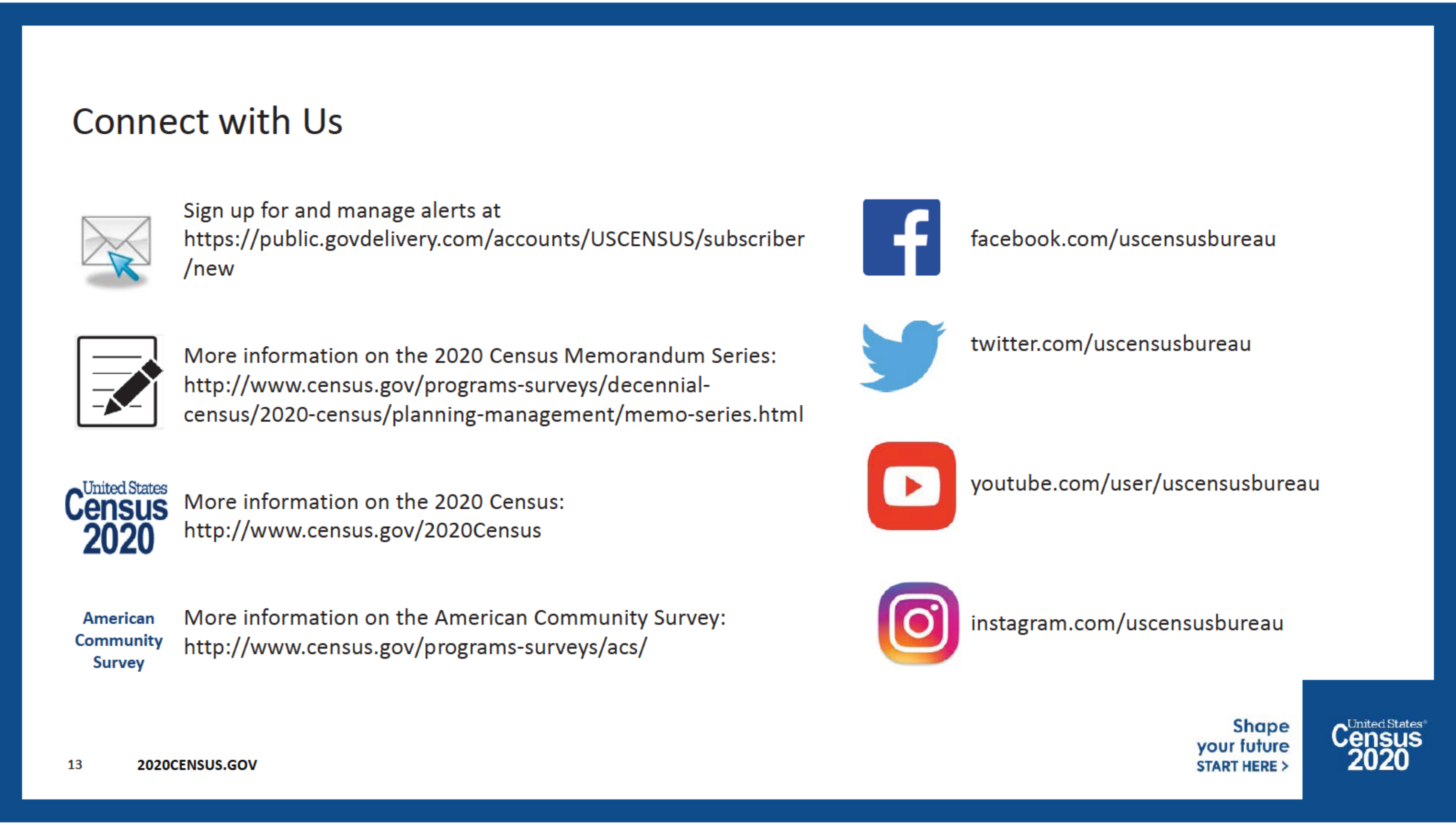
Group Quarters Advance Contact: https://2020census.gov/en/conducting-the-count/gq/gqac.html

eResponse: https://2020census.gov/en/conducting-the-count/gq/eresponse.html

Maritime and Military Vessel Enumeration: https://2020census.gov/en/conducting-the-count/gq/mve.html

Department of Education Student Privacy Policy Office: https://studentprivacy.ed.gov/faq/colleges-and-2020-census

2020 Census Operational Adjustments Due to COVID-19: https://2020census.gov/en/news-events/operational-adjustments-covid-19.html



Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

More information on the 2020 Census: http://www.census.gov/2020Census

More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

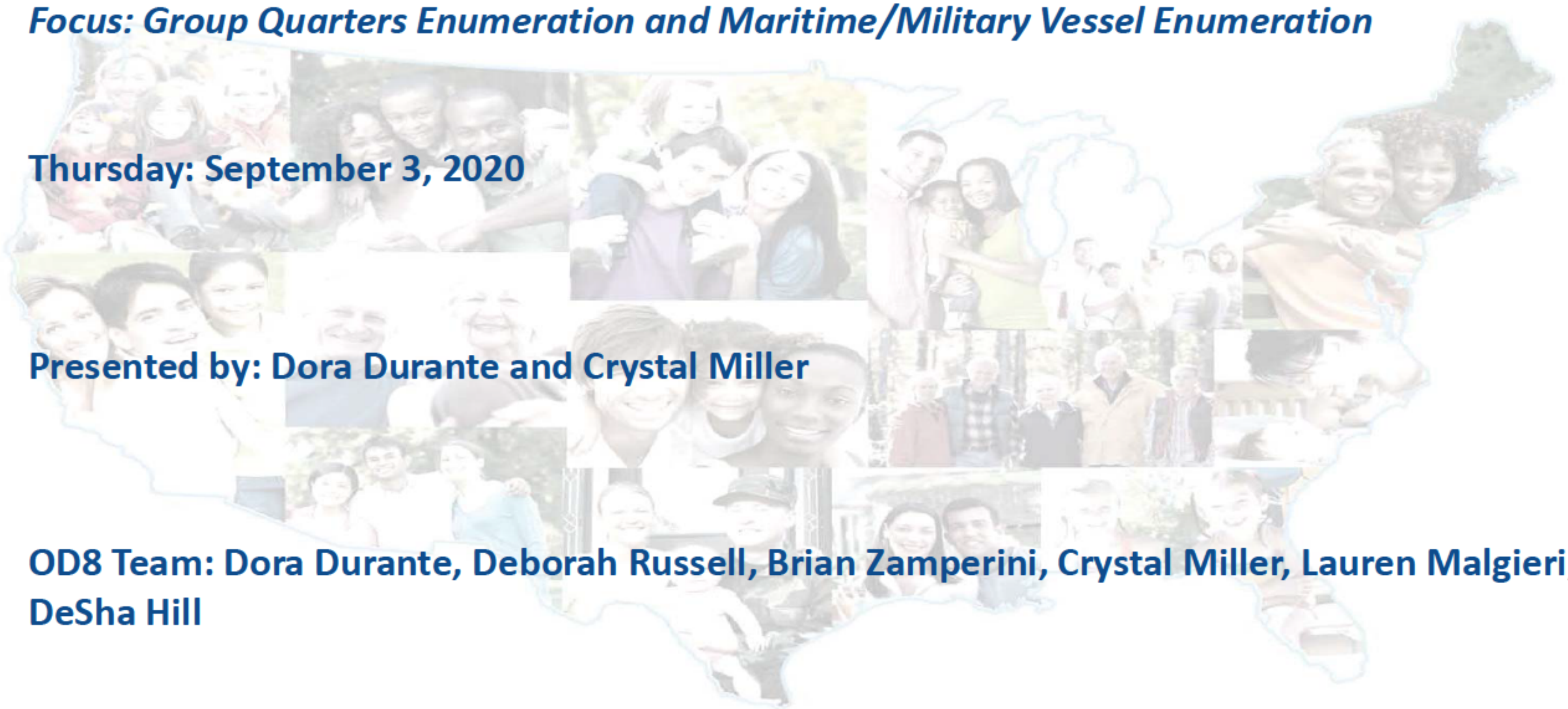
instagram.com/uscensusbureau

United States Census 2020
American Community Survey
Shape your future START HERE >
United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Operational Delivery 8: Group Quarters Update

Focus: Group Quarters Enumeration and Maritime/Military Vessel Enumeration



Thursday: September 3, 2020

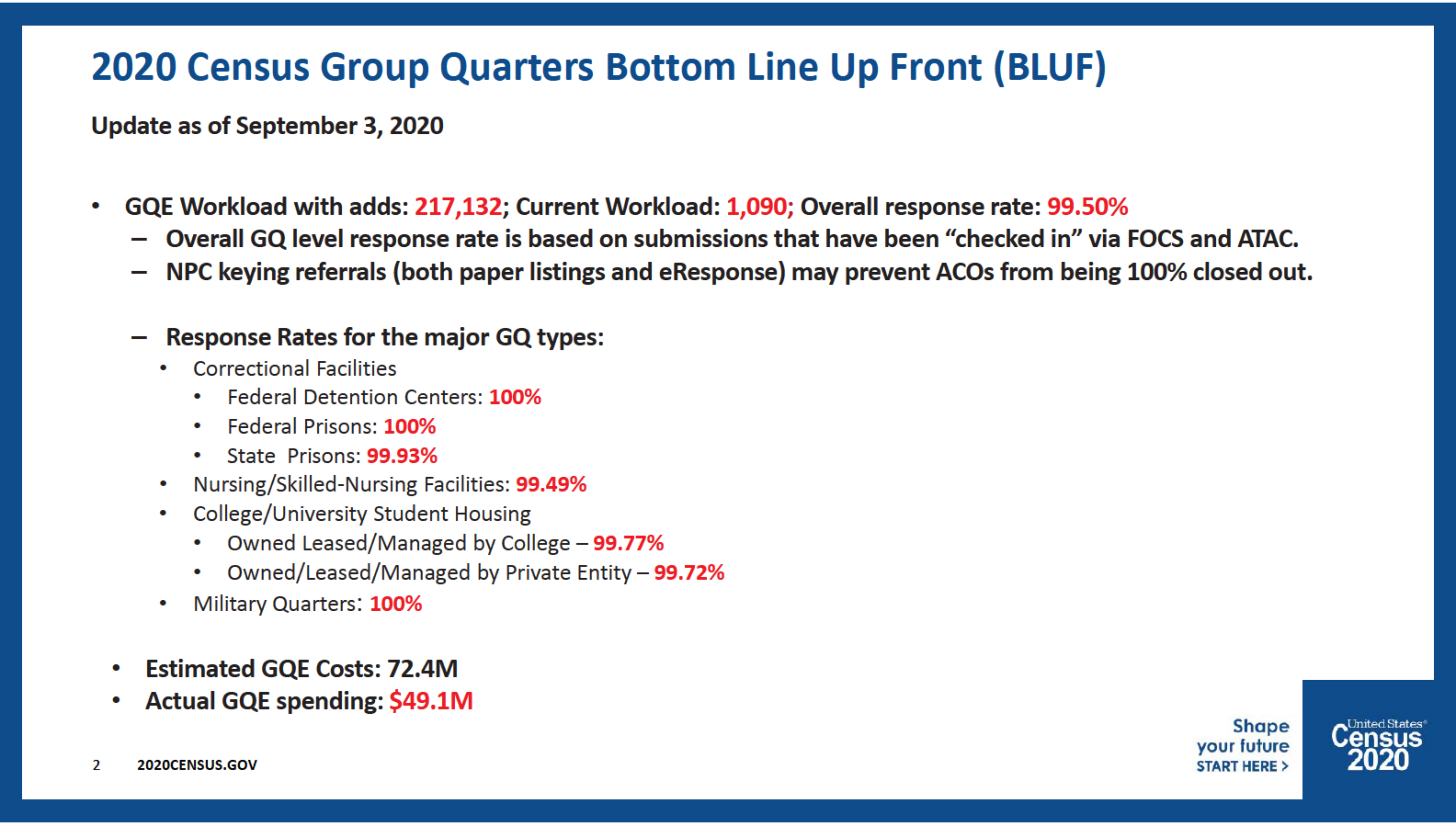
Presented by: Dora Durante and Crystal Miller

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

Shape
your future
START HERE >

United States®
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002



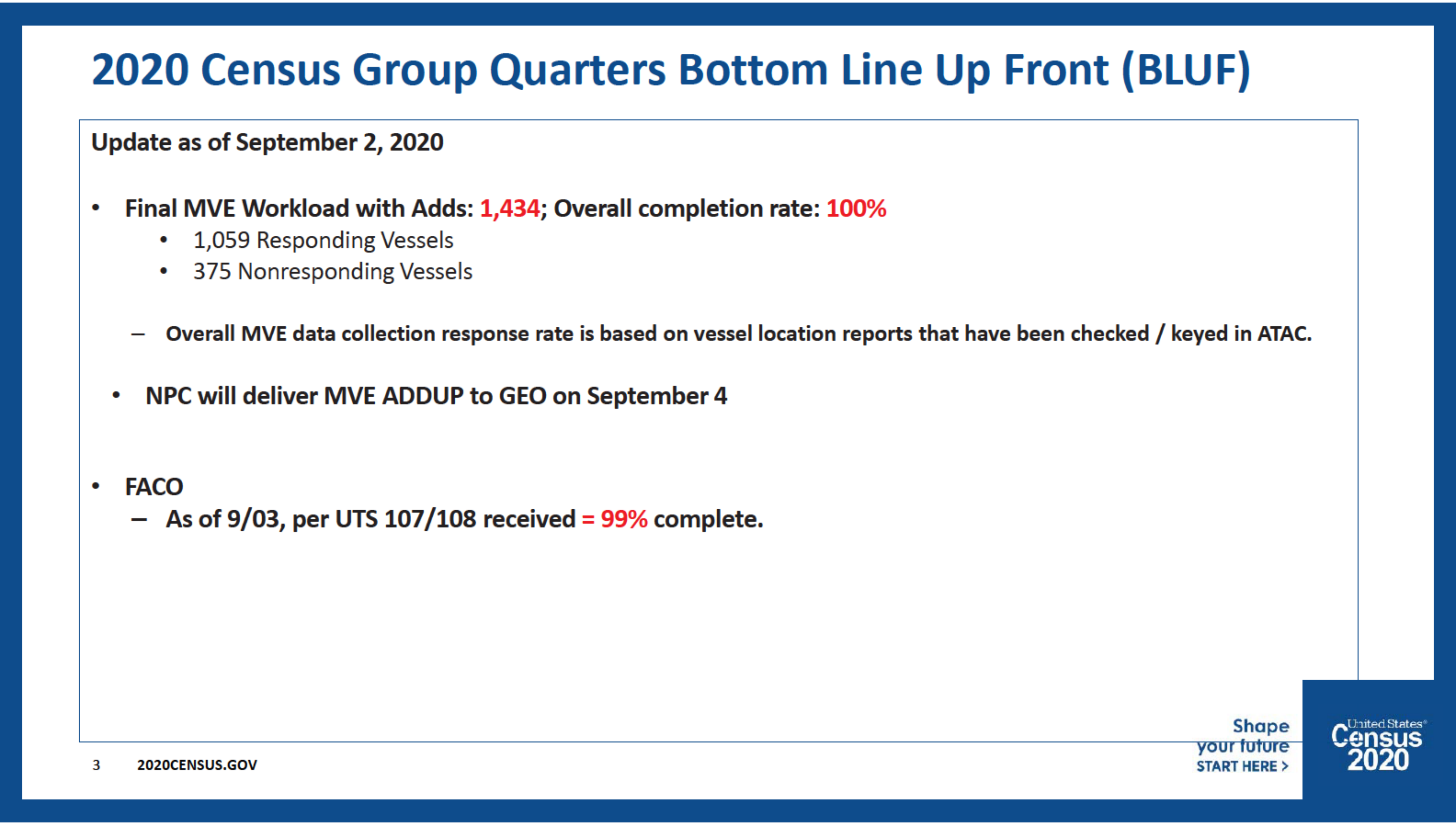
2020 Census Group Quarters Bottom Line Up Front (BLUF)

Update as of September 3, 2020

- **GQE Workload with adds: 217,132; Current Workload: 1,090; Overall response rate: 99.50%**
  - **Overall GQ level response rate is based on submissions that have been "checked in" via FOCS and ATAC.**
  - **NPC keying referrals (both paper listings and eResponse) may prevent ACOs from being 100% closed out.**
  - **Response Rates for the major GQ types:**
    - Correctional Facilities
      - Federal Detention Centers: **100%**
      - Federal Prisons: **100%**
      - State Prisons: **99.93%**
    - Nursing/Skilled-Nursing Facilities: **99.49%**
    - College/University Student Housing
      - Owned Leased/Managed by College – **99.77%**
      - Owned/Leased/Managed by Private Entity – **99.72%**
    - Military Quarters: **100%**

- **Estimated GQE Costs: 72.4M**
- **Actual GQE spending: $49.1M**

DRB Approval Number: CBDRB-FY21-DSEP-002



2020 Census Group Quarters Bottom Line Up Front (BLUF)

Update as of September 2, 2020

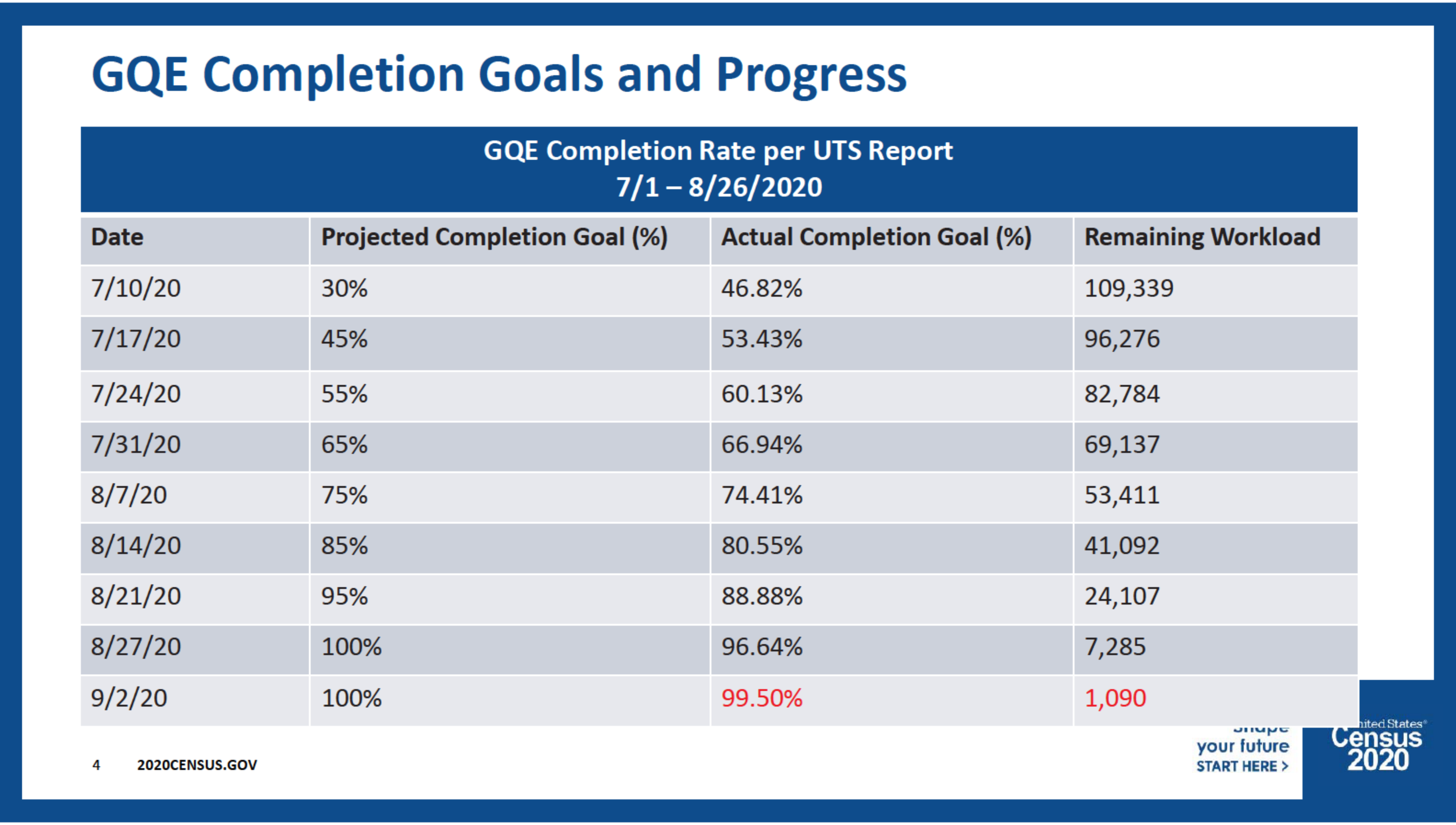
- **Final MVE Workload with Adds: 1,434; Overall completion rate: 100%**
  - 1,059 Responding Vessels
  - 375 Nonresponding Vessels

  – **Overall MVE data collection response rate is based on vessel location reports that have been checked / keyed in ATAC.**

- **NPC will deliver MVE ADDUP to GEO on September 4**

- **FACO**
  – **As of 9/03, per UTS 107/108 received = 99% complete.**

DRB Approval Number: CBDRB-FY21-DSEP-002

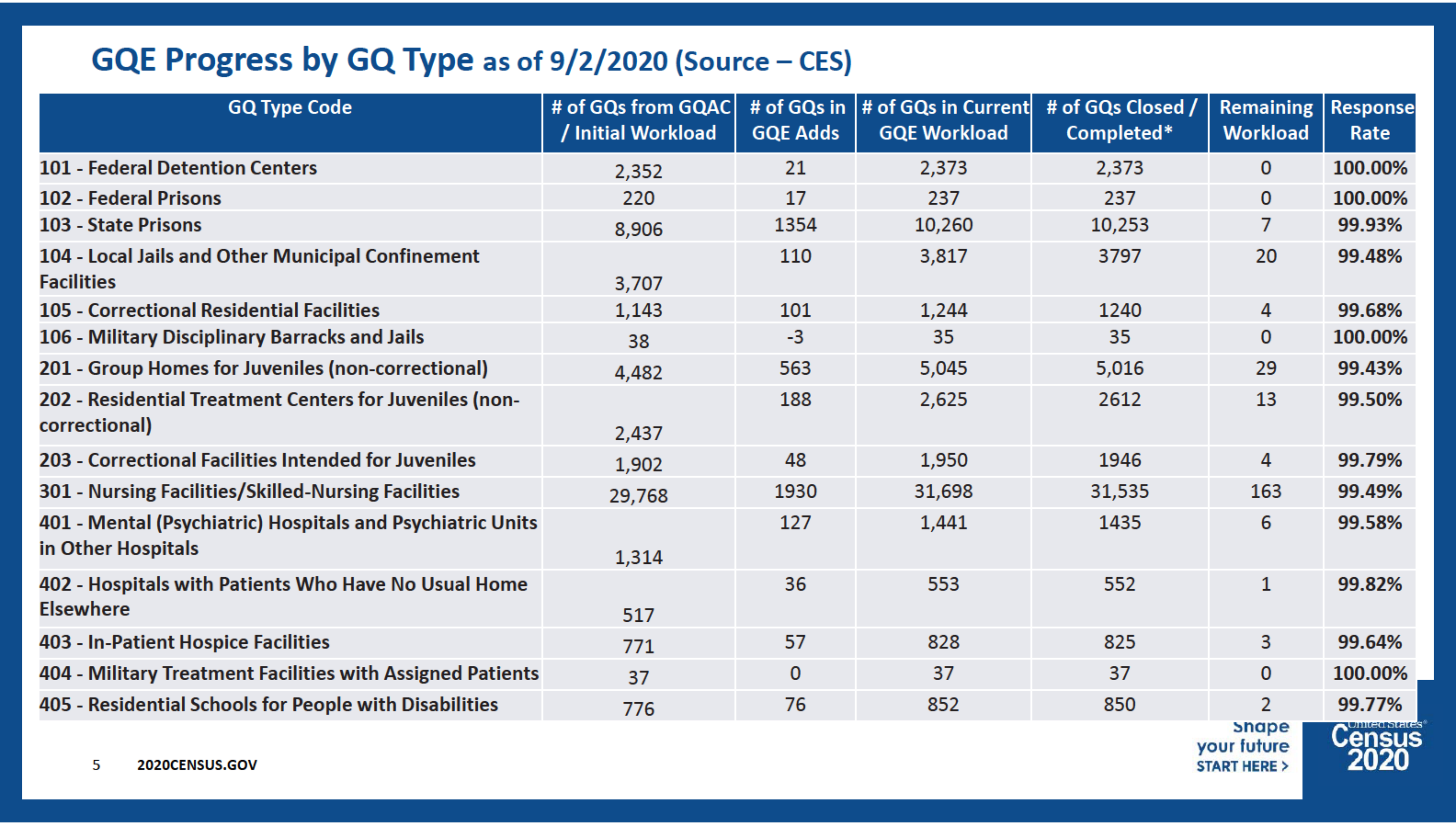


GQE Completion Goals and Progress

GQE Completion Rate per UTS Report
7/1 – 8/26/2020

Date	Projected Completion Goal (%)	Actual Completion Goal (%)	Remaining Workload
7/10/20	30%	46.82%	109,339
7/17/20	45%	53.43%	96,276
7/24/20	55%	60.13%	82,784
7/31/20	65%	66.94%	69,137
8/7/20	75%	74.41%	53,411
8/14/20	85%	80.55%	41,092
8/21/20	95%	88.88%	24,107
8/27/20	100%	96.64%	7,285
9/2/20	100%	99.50%	1,090

DRB Approval Number: CBDRB-FY21-DSEP-002



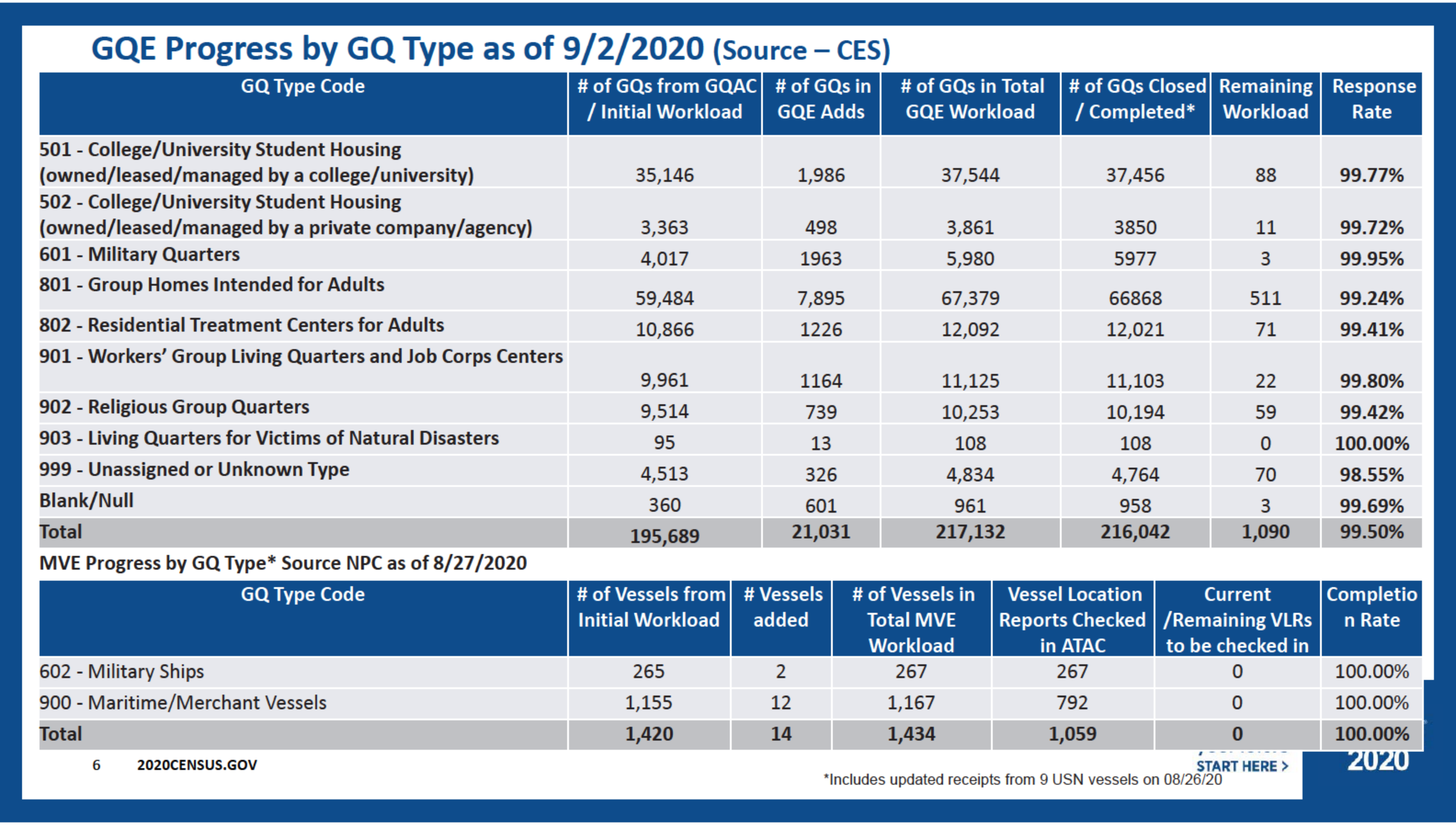
GQE Progress by GQ Type as of 9/2/2020 (Source – CES)

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Current GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
101 - Federal Detention Centers	2,352	21	2,373	2,373	0	100.00%
102 - Federal Prisons	220	17	237	237	0	100.00%
103 - State Prisons	8,906	1354	10,260	10,253	7	99.93%
104 - Local Jails and Other Municipal Confinement Facilities	3,707	110	3,817	3797	20	99.48%
105 - Correctional Residential Facilities	1,143	101	1,244	1240	4	99.68%
106 - Military Disciplinary Barracks and Jails	38	-3	35	35	0	100.00%
201 - Group Homes for Juveniles (non-correctional)	4,482	563	5,045	5,016	29	99.43%
202 - Residential Treatment Centers for Juveniles (non-correctional)	2,437	188	2,625	2612	13	99.50%
203 - Correctional Facilities Intended for Juveniles	1,902	48	1,950	1946	4	99.79%
301 - Nursing Facilities/Skilled-Nursing Facilities	29,768	1930	31,698	31,535	163	99.49%
401 - Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals	1,314	127	1,441	1435	6	99.58%
402 - Hospitals with Patients Who Have No Usual Home Elsewhere	517	36	553	552	1	99.82%
403 - In-Patient Hospice Facilities	771	57	828	825	3	99.64%
404 - Military Treatment Facilities with Assigned Patients	37	0	37	37	0	100.00%
405 - Residential Schools for People with Disabilities	776	76	852	850	2	99.77%

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



GQE Progress by GQ Type as of 9/2/2020 (Source – CES)

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Total GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
501 - College/University Student Housing (owned/leased/managed by a college/university)	35,146	1,986	37,544	37,456	88	**99.77%**
502 - College/University Student Housing (owned/leased/managed by a private company/agency)	3,363	498	3,861	3850	11	**99.72%**
601 - Military Quarters	4,017	1963	5,980	5977	3	**99.95%**
801 - Group Homes Intended for Adults	59,484	7,895	67,379	66868	511	**99.24%**
802 - Residential Treatment Centers for Adults	10,866	1226	12,092	12,021	71	**99.41%**
901 - Workers' Group Living Quarters and Job Corps Centers	9,961	1164	11,125	11,103	22	**99.80%**
902 - Religious Group Quarters	9,514	739	10,253	10,194	59	**99.42%**
903 - Living Quarters for Victims of Natural Disasters	95	13	108	108	0	**100.00%**
999 - Unassigned or Unknown Type	4,513	326	4,834	4,764	70	**98.55%**
Blank/Null	360	601	961	958	3	**99.69%**
Total	**195,689**	**21,031**	**217,132**	**216,042**	**1,090**	**99.50%**

MVE Progress by GQ Type* Source NPC as of 8/27/2020

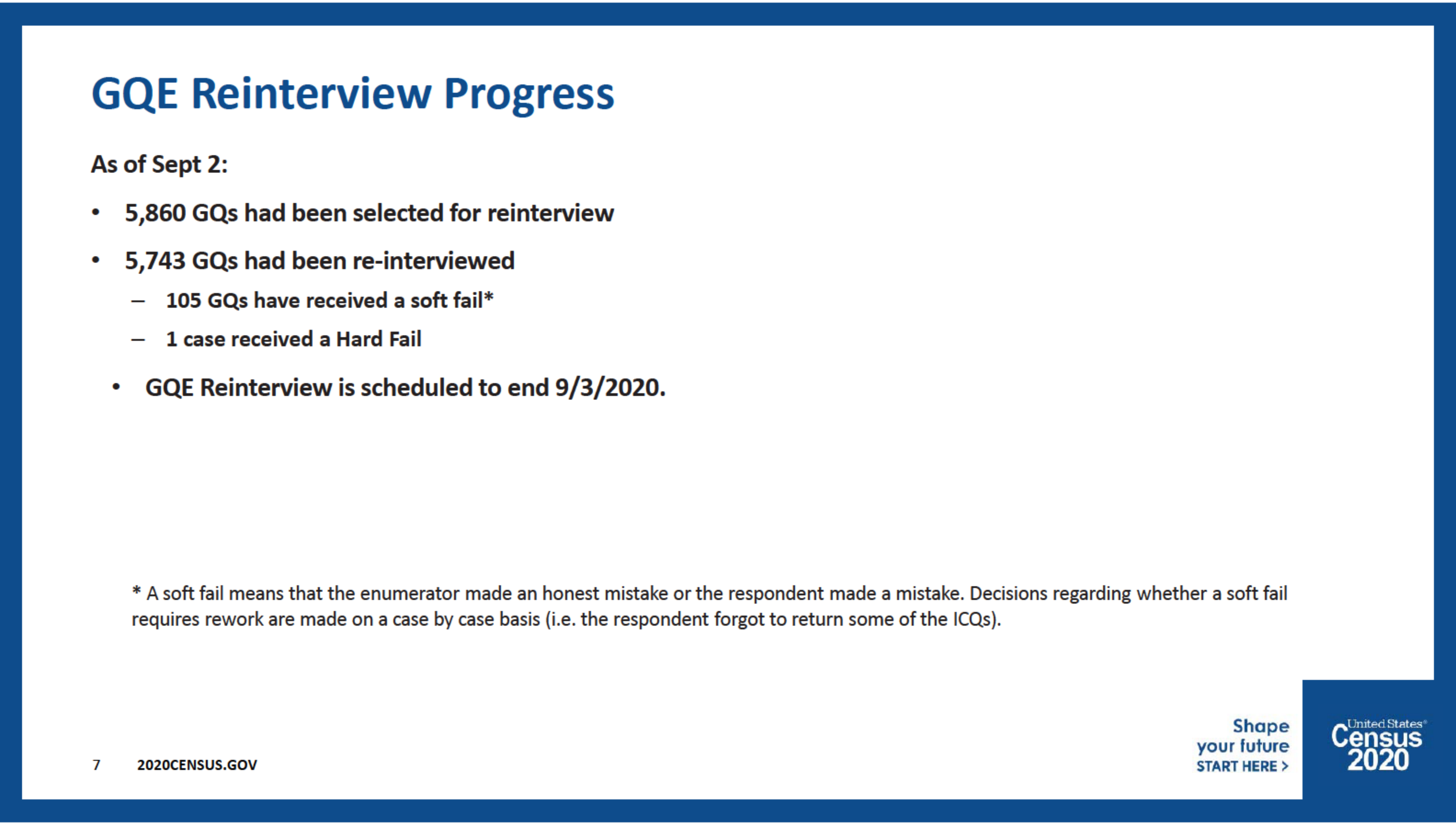
GQ Type Code	# of Vessels from Initial Workload	# Vessels added	# of Vessels in Total MVE Workload	Vessel Location Reports Checked in ATAC	Current /Remaining VLRs to be checked in	Completion Rate
602 - Military Ships	265	2	267	267	0	100.00%
900 - Maritime/Merchant Vessels	1,155	12	1,167	792	0	100.00%
Total	**1,420**	**14**	**1,434**	**1,059**	**0**	**100.00%**

*Includes updated receipts from 9 USN vessels on 08/26/20

6 2020CENSUS.GOV

START HERE >

2020



GQE Reinterview Progress

As of Sept 2:

- **5,860 GQs had been selected for reinterview**
- **5,743 GQs had been re-interviewed**
  - **105 GQs have received a soft fail***
  - **1 case received a Hard Fail**
- **GQE Reinterview is scheduled to end 9/3/2020.**

* A soft fail means that the enumerator made an honest mistake or the respondent made a mistake. Decisions regarding whether a soft fail requires rework are made on a case by case basis (i.e. the respondent forgot to return some of the ICQs).

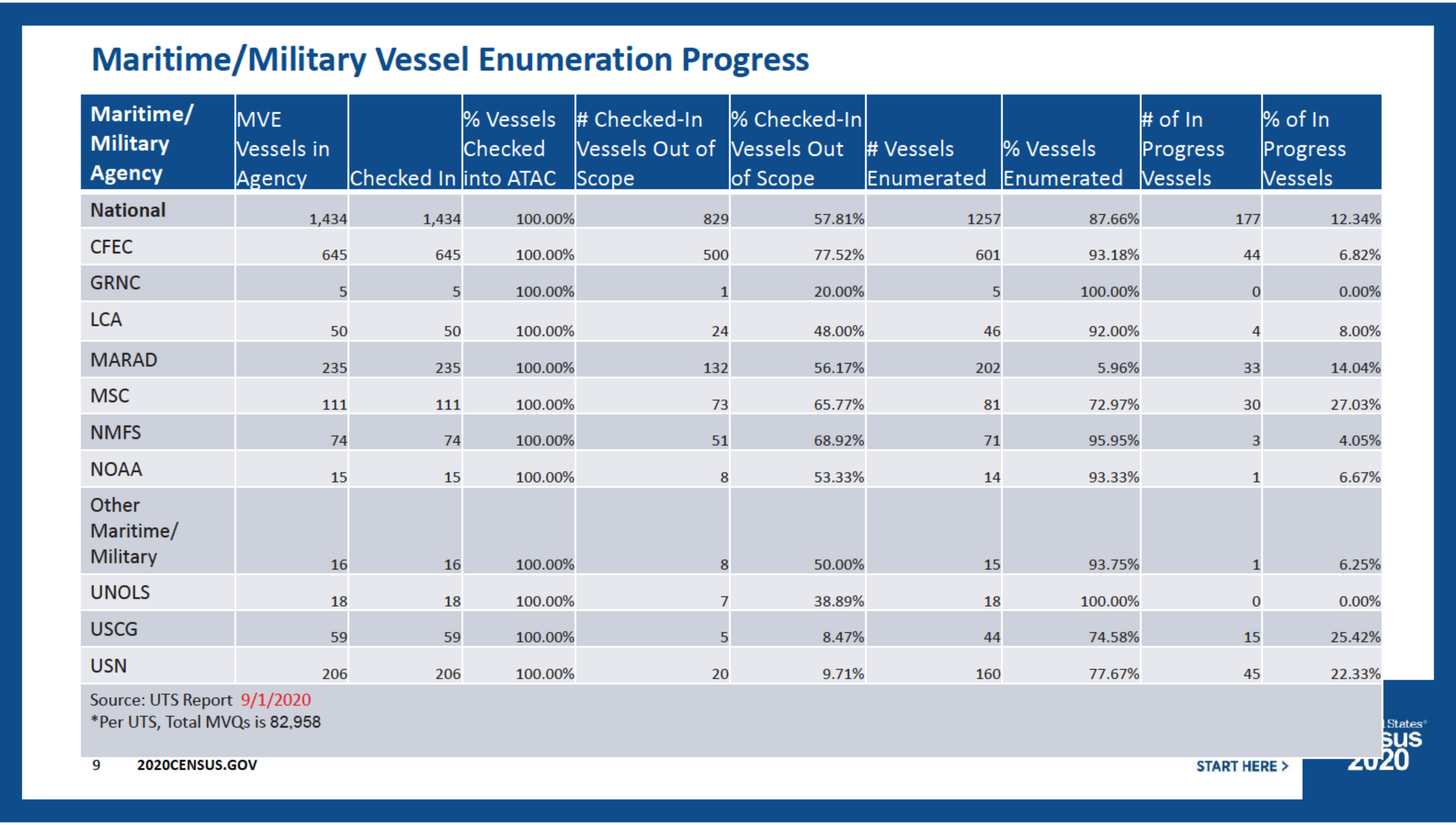
Shape your future START HERE >

United States Census 2020

Backup

Schedules and Systems

DRB Approval Number: CBDRB-FY21-DSEP-002



Maritime/Military Vessel Enumeration Progress

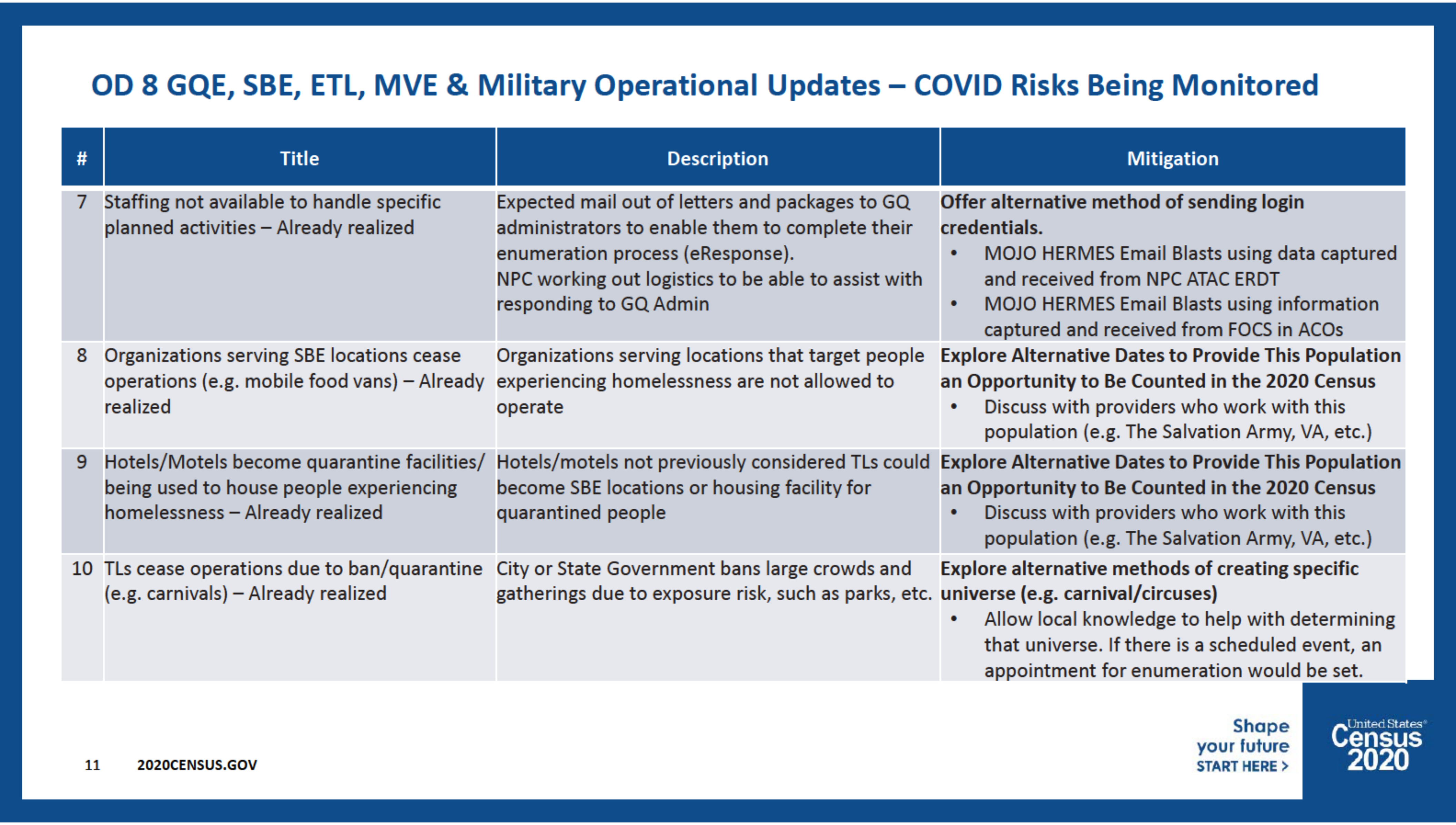
Maritime/ Military Agency	MVE Vessels in Agency	Checked In	% Vessels Checked into ATAC	# Checked-In Vessels Out of Scope	% Checked-In Vessels Out of Scope	# Vessels Enumerated	% Vessels Enumerated	# of In Progress Vessels	% of In Progress Vessels
National	1,434	1,434	100.00%	829	57.81%	1257	87.66%	177	12.34%
CFEC	645	645	100.00%	500	77.52%	601	93.18%	44	6.82%
GRNC	5	5	100.00%	1	20.00%	5	100.00%	0	0.00%
LCA	50	50	100.00%	24	48.00%	46	92.00%	4	8.00%
MARAD	235	235	100.00%	132	56.17%	202	5.96%	33	14.04%
MSC	111	111	100.00%	73	65.77%	81	72.97%	30	27.03%
NMFS	74	74	100.00%	51	68.92%	71	95.95%	3	4.05%
NOAA	15	15	100.00%	8	53.33%	14	93.33%	1	6.67%
Other Maritime/ Military	16	16	100.00%	8	50.00%	15	93.75%	1	6.25%
UNOLS	18	18	100.00%	7	38.89%	18	100.00%	0	0.00%
USCG	59	59	100.00%	5	8.47%	44	74.58%	15	25.42%
USN	206	206	100.00%	20	9.71%	160	77.67%	45	22.33%

Source: UTS Report 9/1/2020
*Per UTS, Total MVQs is 82,958

START HERE >

OD 8 GQE Milestone Conduct & Training Dates

Conduct Activity	Proposed Start	Proposed Finish
Conduct MVE Operation	04/01/20 (A)	08/21/20 (P)
Conduct GQE eResponse Operation	04/01/20 (A)	08/26/20 (A)
Conduct GQE Operation Field	04/20/20 (A)	08/26/20 (P)
Conduct GQE In-Person Operation	07/01/20 (A)	08/26/20 (P)
Conduct GQ Reinterview Operation	07/02/20 (A)	09/03/20 (P)
Conduct DVS Enumeration Field	07/06/20 (A)	08/26/20 (P)

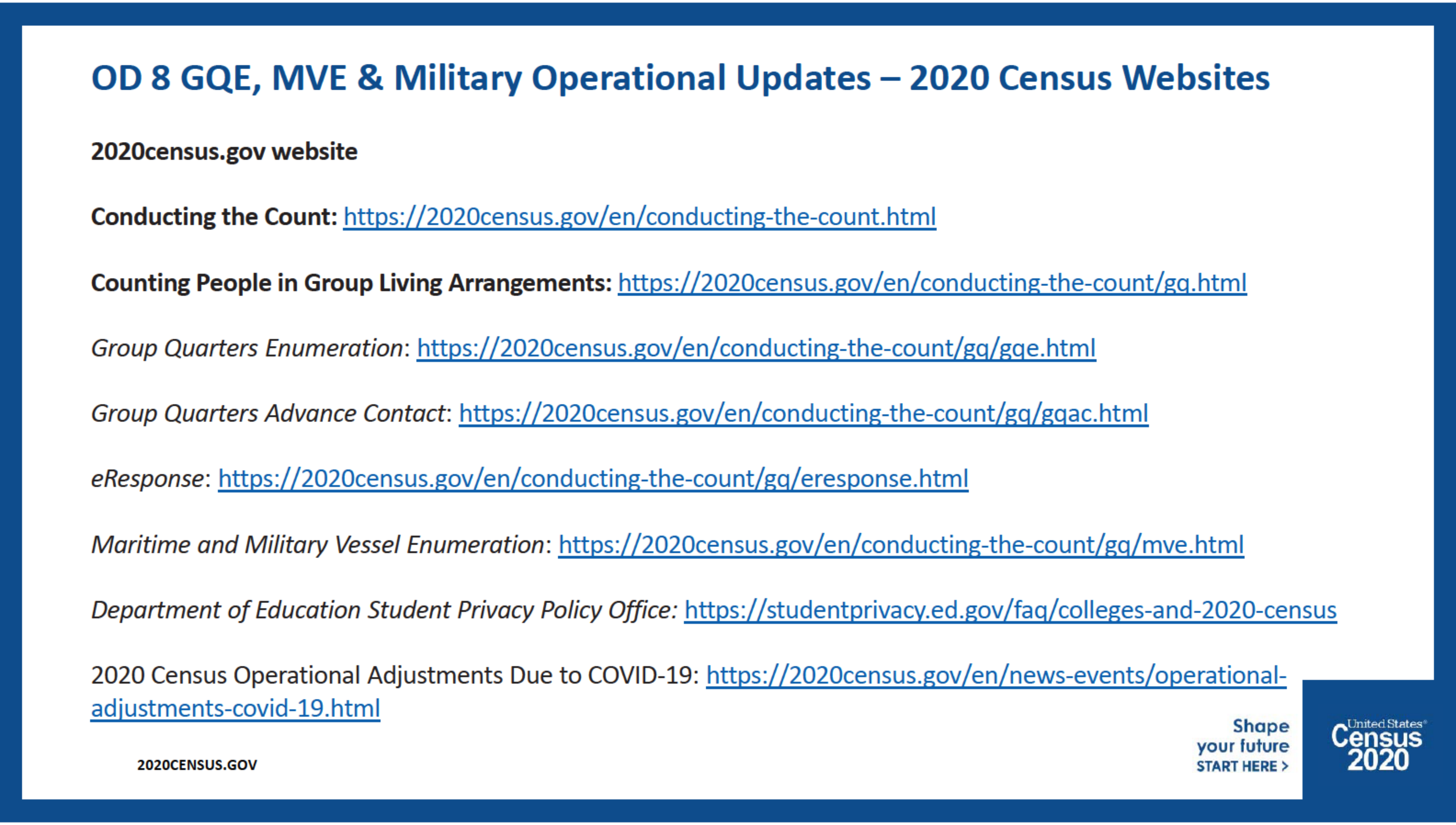


OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Risks Being Monitored

#	Title	Description	Mitigation
7	Staffing not available to handle specific planned activities – Already realized	Expected mail out of letters and packages to GQ administrators to enable them to complete their enumeration process (eResponse). NPC working out logistics to be able to assist with responding to GQ Admin	**Offer alternative method of sending login credentials.** • MOJO HERMES Email Blasts using data captured and received from NPC ATAC ERDT • MOJO HERMES Email Blasts using information captured and received from FOCS in ACOs
8	Organizations serving SBE locations cease operations (e.g. mobile food vans) – Already realized	Organizations serving locations that target people experiencing homelessness are not allowed to operate	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
9	Hotels/Motels become quarantine facilities/ being used to house people experiencing homelessness – Already realized	Hotels/motels not previously considered TLs could become SBE locations or housing facility for quarantined people	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
10	TLs cease operations due to ban/quarantine (e.g. carnivals) – Already realized	City or State Government bans large crowds and gatherings due to exposure risk, such as parks, etc.	**Explore alternative methods of creating specific universe (e.g. carnival/circuses)** • Allow local knowledge to help with determining that universe. If there is a scheduled event, an appointment for enumeration would be set.

Shape
your future
START HERE >

United States Census 2020



OD 8 GQE, MVE & Military Operational Updates – 2020 Census Websites

2020census.gov website

Conducting the Count: https://2020census.gov/en/conducting-the-count.html

Counting People in Group Living Arrangements: https://2020census.gov/en/conducting-the-count/gq.html

Group Quarters Enumeration: https://2020census.gov/en/conducting-the-count/gq/gqe.html

Group Quarters Advance Contact: https://2020census.gov/en/conducting-the-count/gq/gqac.html

eResponse: https://2020census.gov/en/conducting-the-count/gq/eresponse.html

Maritime and Military Vessel Enumeration: https://2020census.gov/en/conducting-the-count/gq/mve.html

Department of Education Student Privacy Policy Office: https://studentprivacy.ed.gov/faq/colleges-and-2020-census

2020 Census Operational Adjustments Due to COVID-19: https://2020census.gov/en/news-events/operational-adjustments-covid-19.html

2020CENSUS.GOV

Shape
your future
START HERE >

United States®
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002

Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

United States Census 2020

More information on the 2020 Census: http://www.census.gov/2020Census

American Community Survey

More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

instagram.com/uscensusbureau

Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

2020 Census Operational Delivery 8: Group Quarters Update

Focus: Group Quarters Enumeration and Maritime/Military Vessel Enumeration

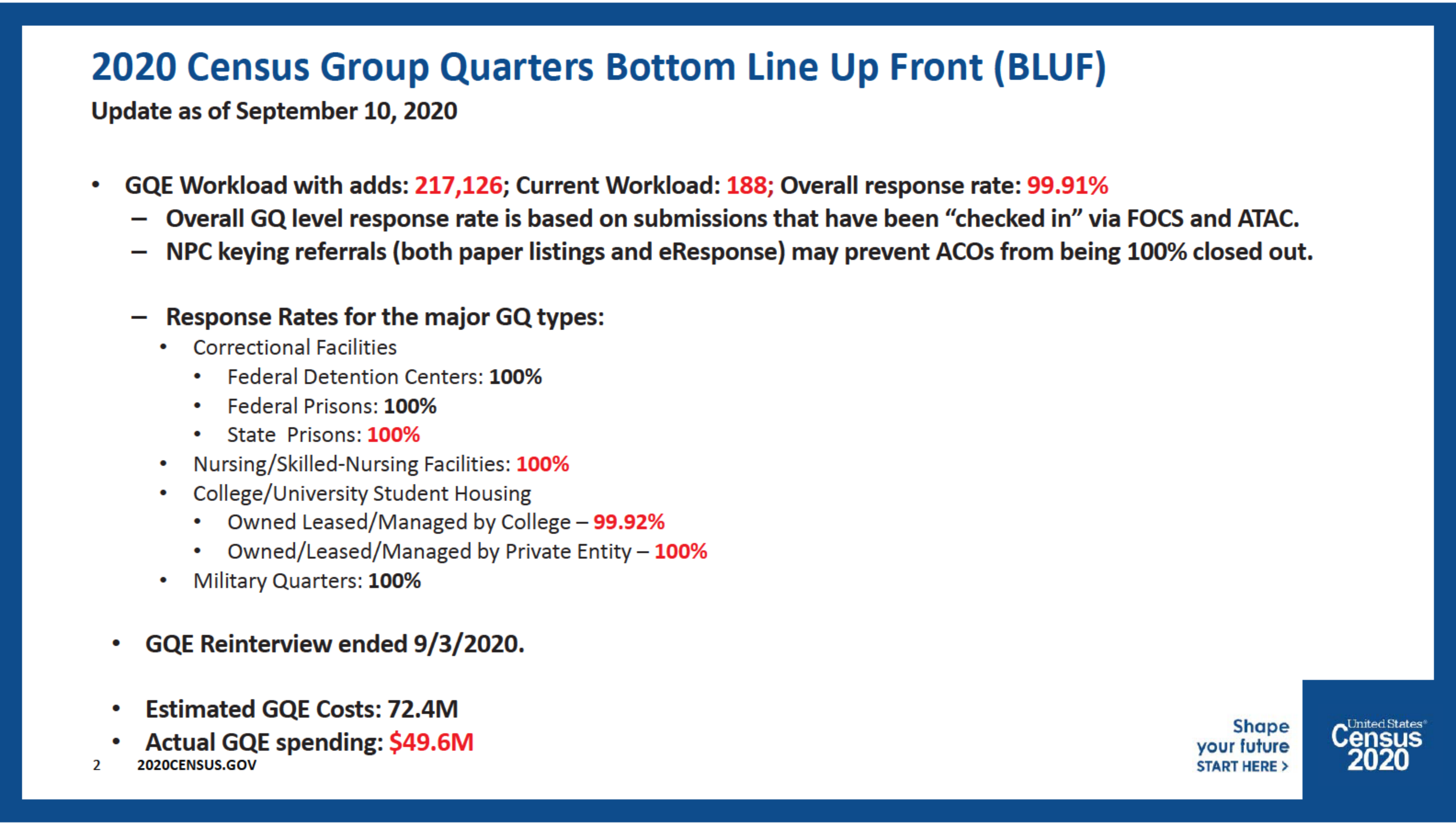
Thursday: September 10, 2020

Presented by: Dora Durante and Crystal Miller

OD8 Team: Dora Durante, Deborah Russell, Brian Zamperini, Crystal Miller, Lauren Malgieri, Sonya DeSha Hill

Shape
your future
START HERE >

United States®
Census
2020



2020 Census Group Quarters Bottom Line Up Front (BLUF)

Update as of September 10, 2020

- **GQE Workload with adds: 217,126; Current Workload: 188; Overall response rate: 99.91%**
  - **Overall GQ level response rate is based on submissions that have been "checked in" via FOCS and ATAC.**
  - **NPC keying referrals (both paper listings and eResponse) may prevent ACOs from being 100% closed out.**
  - **Response Rates for the major GQ types:**
    - Correctional Facilities
      - Federal Detention Centers: **100%**
      - Federal Prisons: **100%**
      - State Prisons: **100%**
    - Nursing/Skilled-Nursing Facilities: **100%**
    - College/University Student Housing
      - Owned Leased/Managed by College – **99.92%**
      - Owned/Leased/Managed by Private Entity – **100%**
    - Military Quarters: **100%**

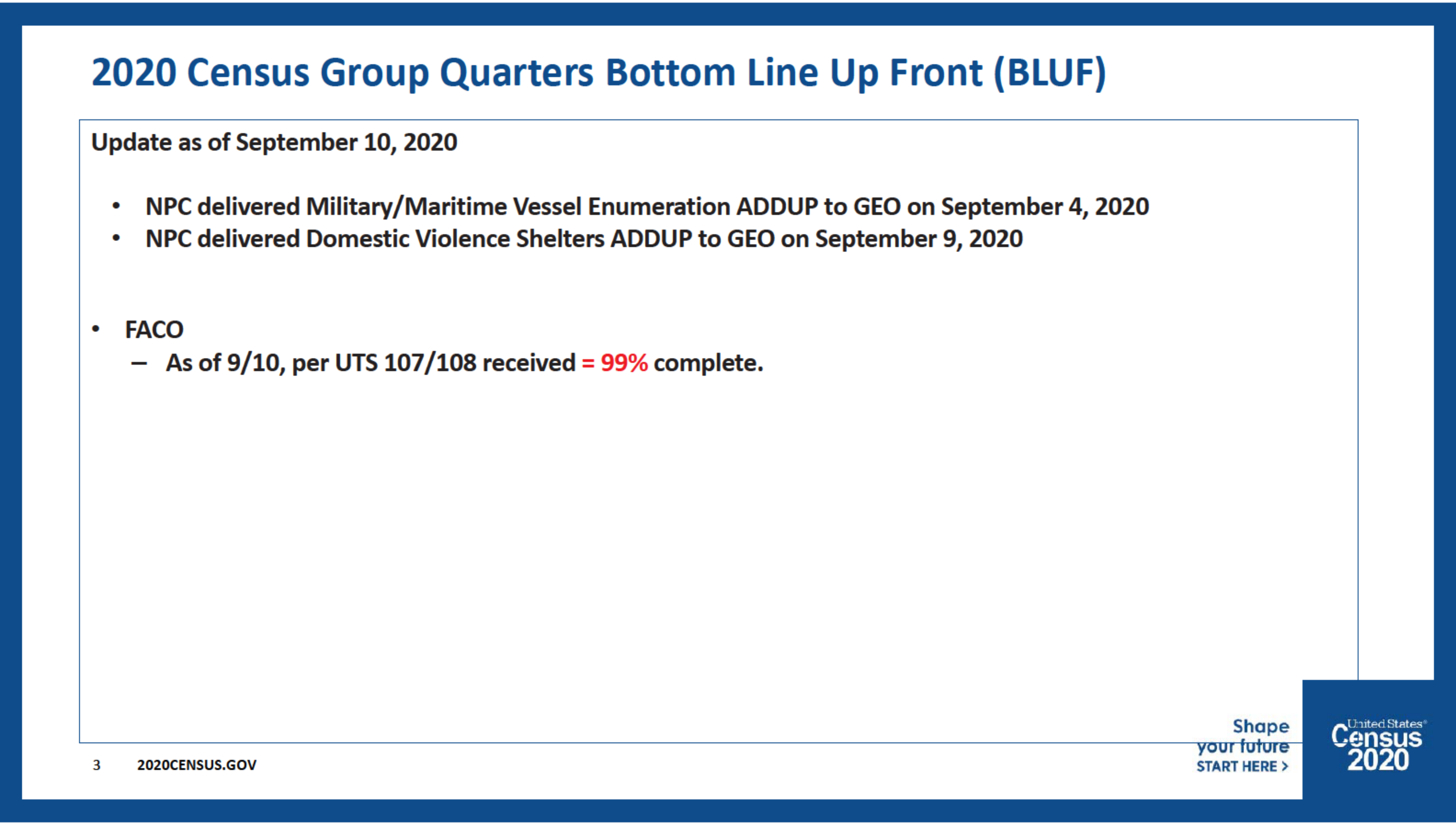
- **GQE Reinterview ended 9/3/2020.**

- **Estimated GQE Costs: 72.4M**
- **Actual GQE spending: $49.6M**

Shape
your future
START HERE >

United States®
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002



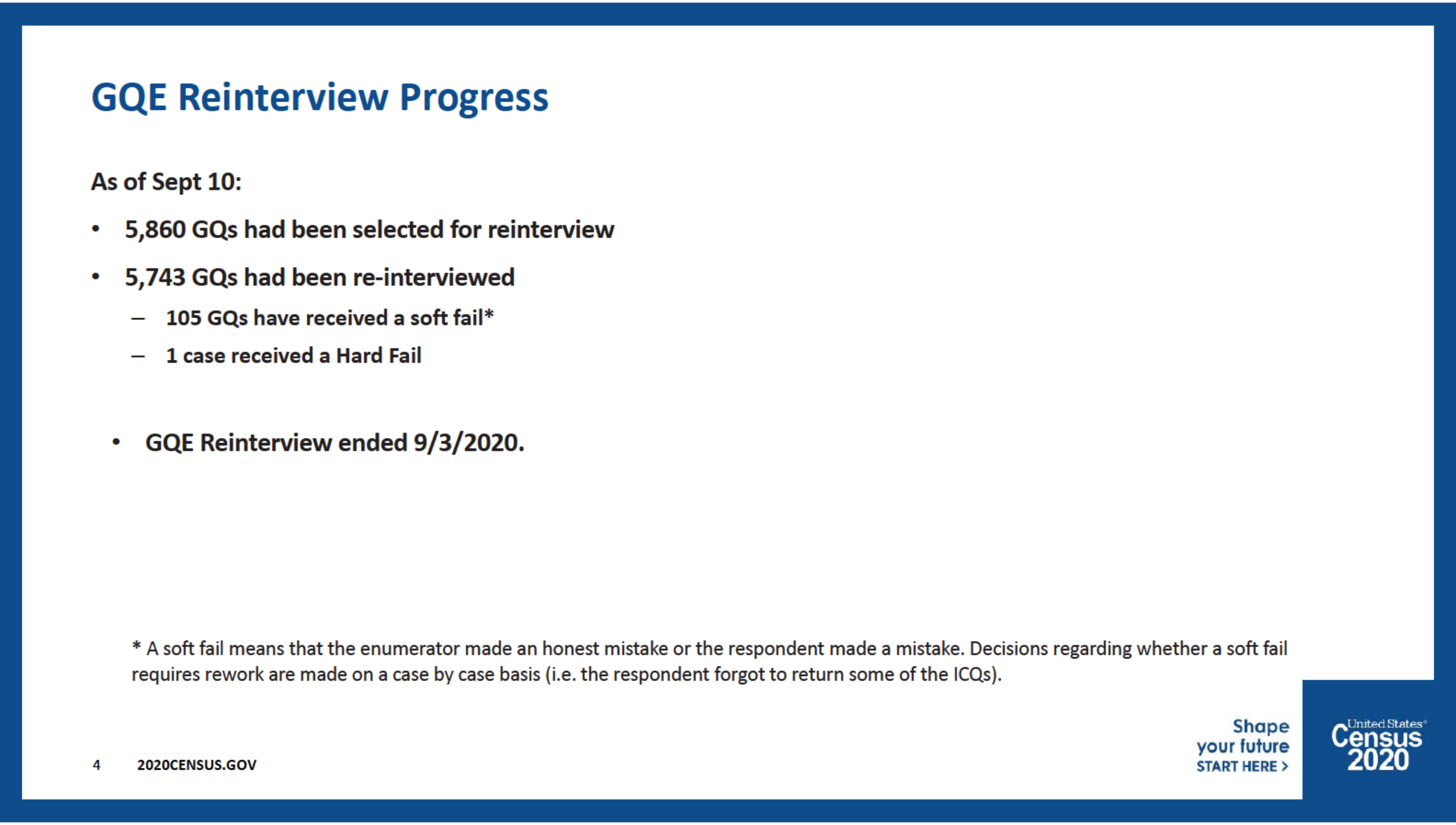
2020 Census Group Quarters Bottom Line Up Front (BLUF)

Update as of September 10, 2020

- **NPC delivered Military/Maritime Vessel Enumeration ADDUP to GEO on September 4, 2020**
- **NPC delivered Domestic Violence Shelters ADDUP to GEO on September 9, 2020**

- **FACO**
  - **As of 9/10, per UTS 107/108 received = 99% complete.**

DRB Approval Number: CBDRB-FY21-DSEP-002



GQE Reinterview Progress

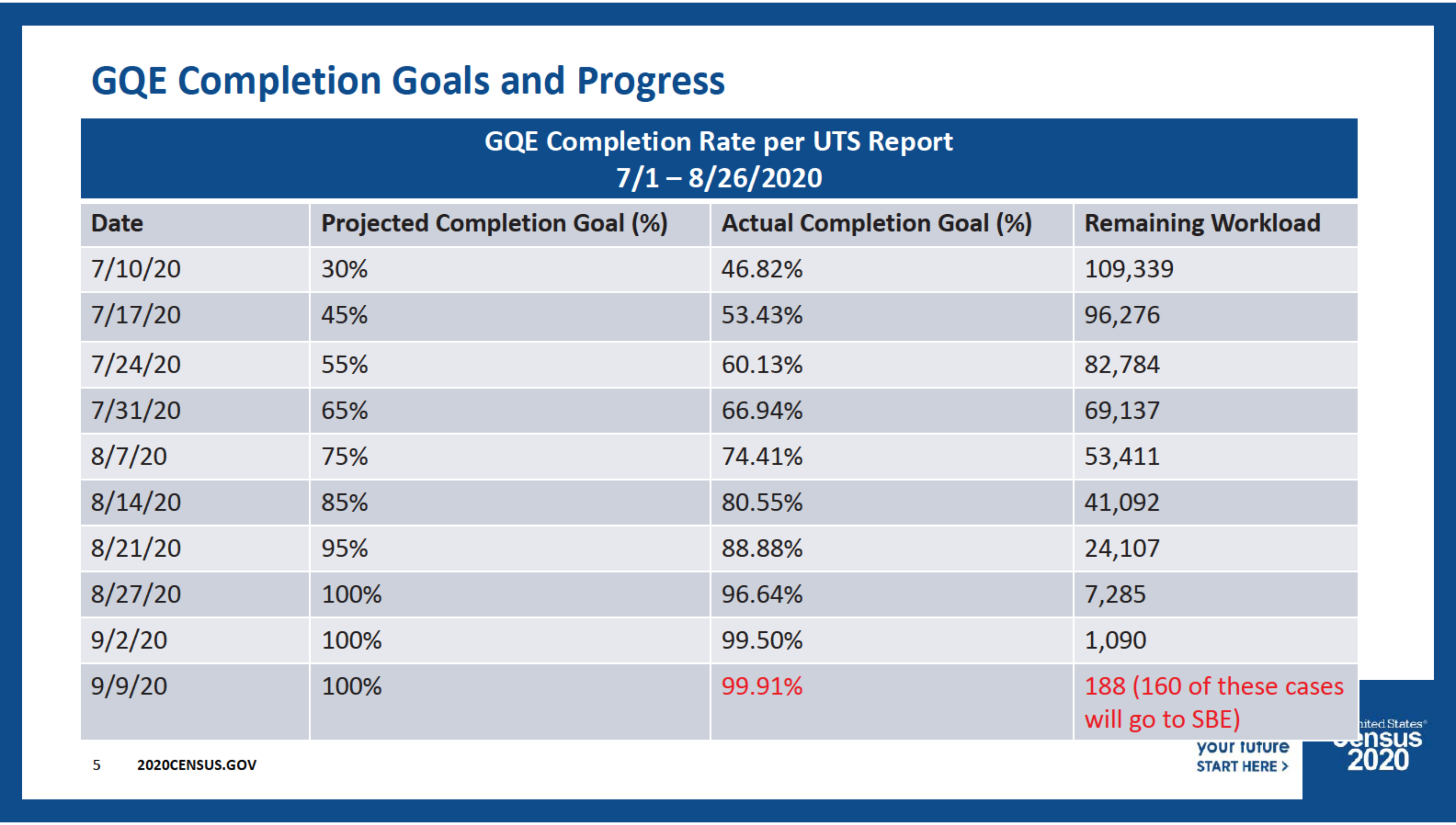
As of Sept 10:

- **5,860 GQs had been selected for reinterview**
- **5,743 GQs had been re-interviewed**
  - **105 GQs have received a soft fail***
  - **1 case received a Hard Fail**

- **GQE Reinterview ended 9/3/2020.**

* A soft fail means that the enumerator made an honest mistake or the respondent made a mistake. Decisions regarding whether a soft fail requires rework are made on a case by case basis (i.e. the respondent forgot to return some of the ICQs).

DRB Approval Number: CBDRB-FY21-DSEP-002

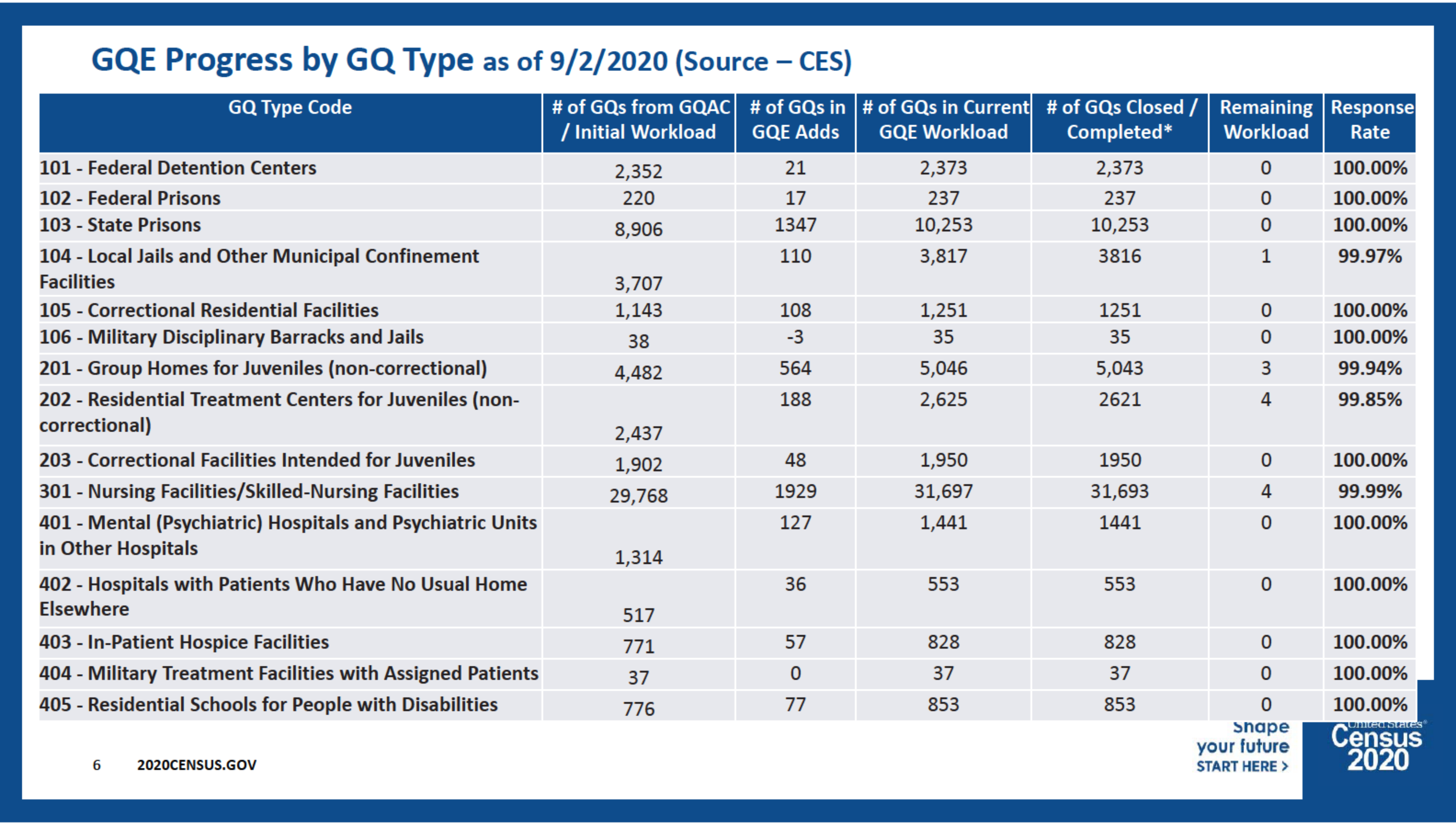


GQE Completion Goals and Progress

GQE Completion Rate per UTS Report
7/1 – 8/26/2020

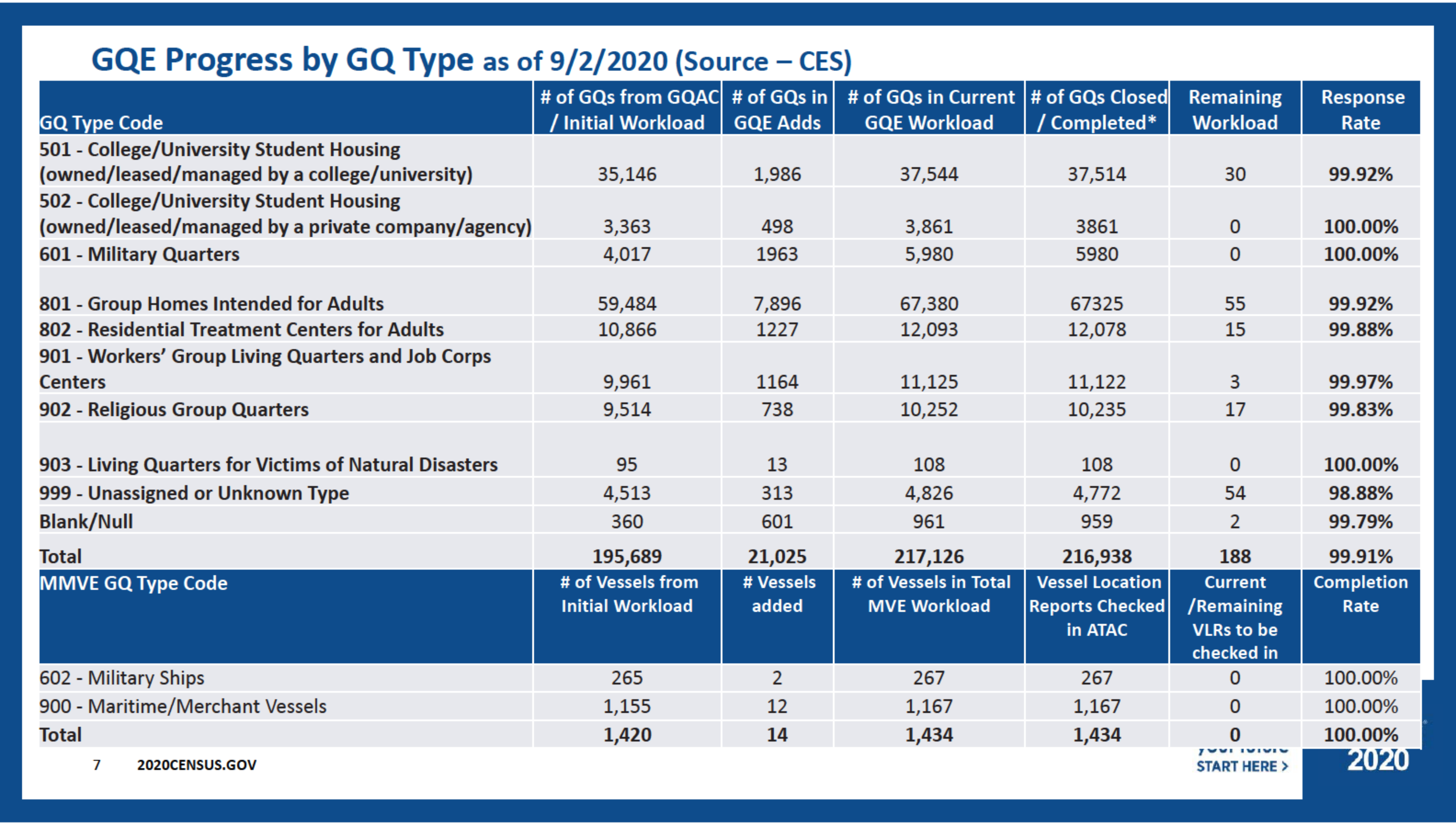
Date	Projected Completion Goal (%)	Actual Completion Goal (%)	Remaining Workload
7/10/20	30%	46.82%	109,339
7/17/20	45%	53.43%	96,276
7/24/20	55%	60.13%	82,784
7/31/20	65%	66.94%	69,137
8/7/20	75%	74.41%	53,411
8/14/20	85%	80.55%	41,092
8/21/20	95%	88.88%	24,107
8/27/20	100%	96.64%	7,285
9/2/20	100%	99.50%	1,090
9/9/20	100%	99.91%	188 (160 of these cases will go to SBE)

DRB Approval Number: CBDRB-FY21-DSEP-002



GQE Progress by GQ Type as of 9/2/2020 (Source – CES)

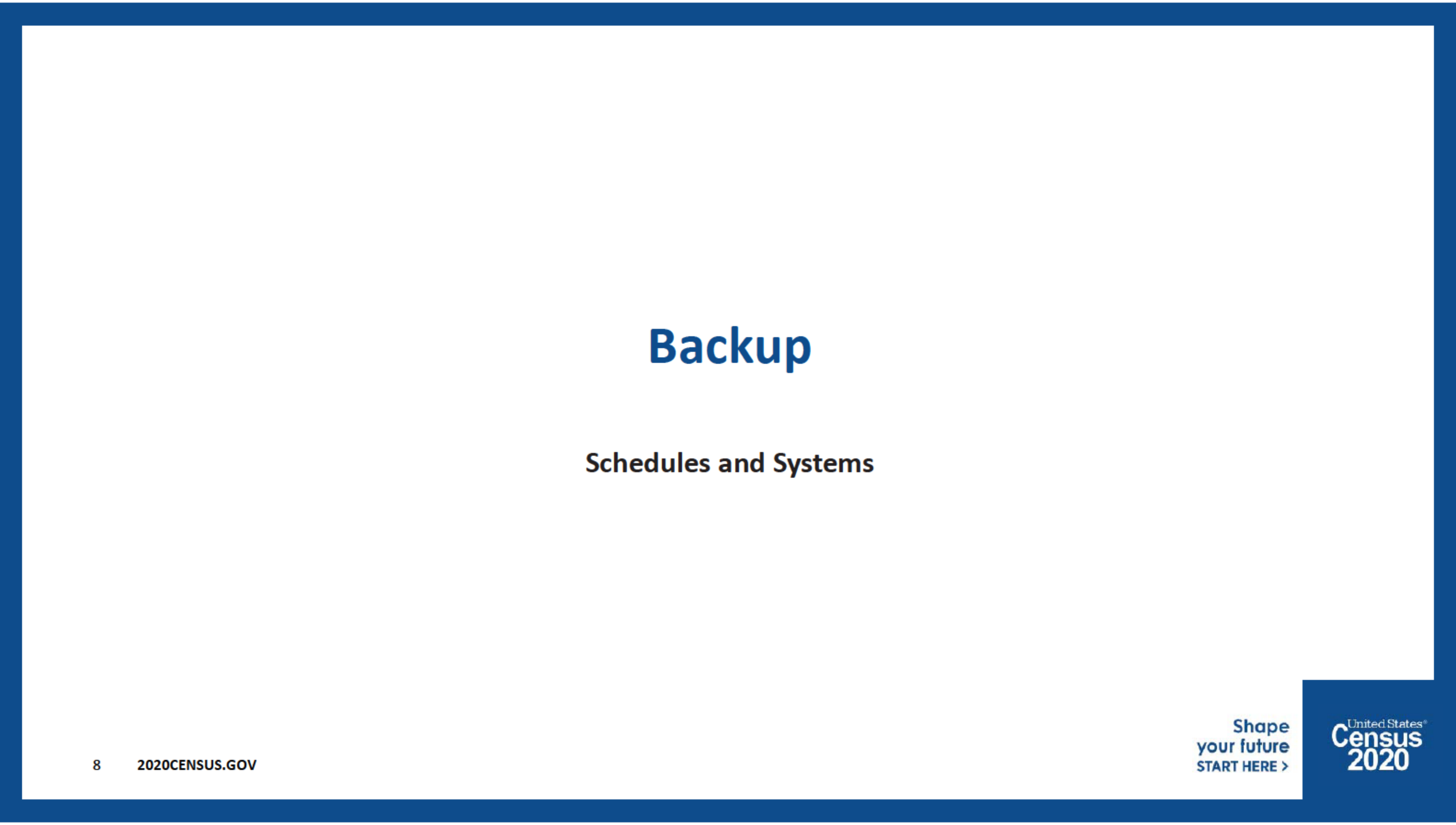
GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Current GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
101 - Federal Detention Centers	2,352	21	2,373	2,373	0	100.00%
102 - Federal Prisons	220	17	237	237	0	100.00%
103 - State Prisons	8,906	1347	10,253	10,253	0	100.00%
104 - Local Jails and Other Municipal Confinement Facilities	3,707	110	3,817	3816	1	99.97%
105 - Correctional Residential Facilities	1,143	108	1,251	1251	0	100.00%
106 - Military Disciplinary Barracks and Jails	38	-3	35	35	0	100.00%
201 - Group Homes for Juveniles (non-correctional)	4,482	564	5,046	5,043	3	99.94%
202 - Residential Treatment Centers for Juveniles (non-correctional)	2,437	188	2,625	2621	4	99.85%
203 - Correctional Facilities Intended for Juveniles	1,902	48	1,950	1950	0	100.00%
301 - Nursing Facilities/Skilled-Nursing Facilities	29,768	1929	31,697	31,693	4	99.99%
401 - Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals	1,314	127	1,441	1441	0	100.00%
402 - Hospitals with Patients Who Have No Usual Home Elsewhere	517	36	553	553	0	100.00%
403 - In-Patient Hospice Facilities	771	57	828	828	0	100.00%
404 - Military Treatment Facilities with Assigned Patients	37	0	37	37	0	100.00%
405 - Residential Schools for People with Disabilities	776	77	853	853	0	100.00%



GQE Progress by GQ Type as of 9/2/2020 (Source – CES)

GQ Type Code	# of GQs from GQAC / Initial Workload	# of GQs in GQE Adds	# of GQs in Current GQE Workload	# of GQs Closed / Completed*	Remaining Workload	Response Rate
501 - College/University Student Housing (owned/leased/managed by a college/university)	35,146	1,986	37,544	37,514	30	99.92%
502 - College/University Student Housing (owned/leased/managed by a private company/agency)	3,363	498	3,861	3861	0	100.00%
601 - Military Quarters	4,017	1963	5,980	5980	0	100.00%
801 - Group Homes Intended for Adults	59,484	7,896	67,380	67325	55	99.92%
802 - Residential Treatment Centers for Adults	10,866	1227	12,093	12,078	15	99.88%
901 - Workers' Group Living Quarters and Job Corps Centers	9,961	1164	11,125	11,122	3	99.97%
902 - Religious Group Quarters	9,514	738	10,252	10,235	17	99.83%
903 - Living Quarters for Victims of Natural Disasters	95	13	108	108	0	100.00%
999 - Unassigned or Unknown Type	4,513	313	4,826	4,772	54	98.88%
Blank/Null	360	601	961	959	2	99.79%
Total	**195,689**	**21,025**	**217,126**	**216,938**	**188**	**99.91%**
MMVE GQ Type Code	**# of Vessels from Initial Workload**	**# Vessels added**	**# of Vessels in Total MVE Workload**	**Vessel Location Reports Checked in ATAC**	**Current /Remaining VLRs to be checked in**	**Completion Rate**
602 - Military Ships	265	2	267	267	0	100.00%
900 - Maritime/Merchant Vessels	1,155	12	1,167	1,167	0	100.00%
Total	**1,420**	**14**	**1,434**	**1,434**	**0**	**100.00%**

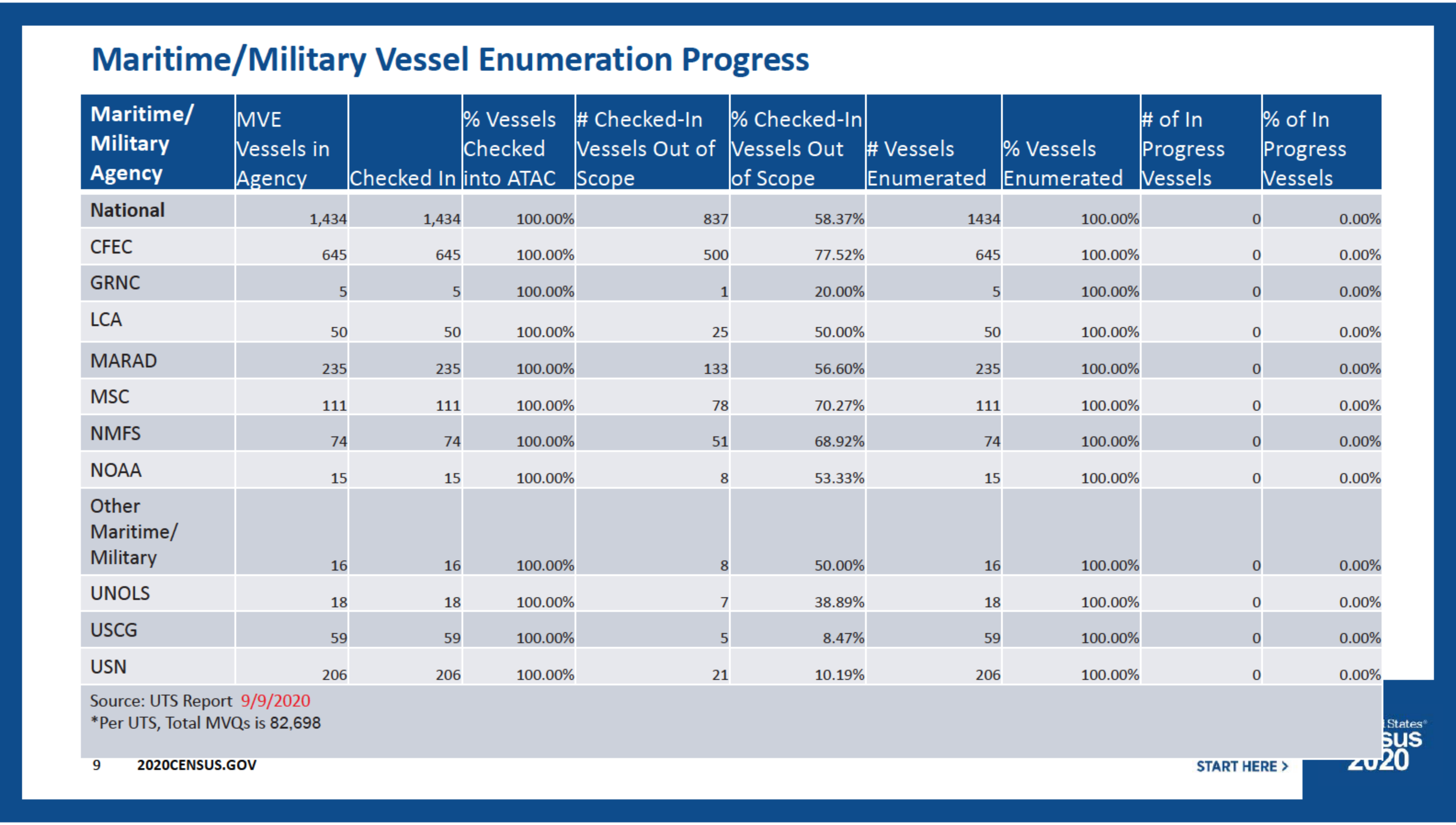
DRB Approval Number: CBDRB-FY21-DSEP-002



Backup

Schedules and Systems

DRB Approval Number: CBDRB-FY21-DSEP-002

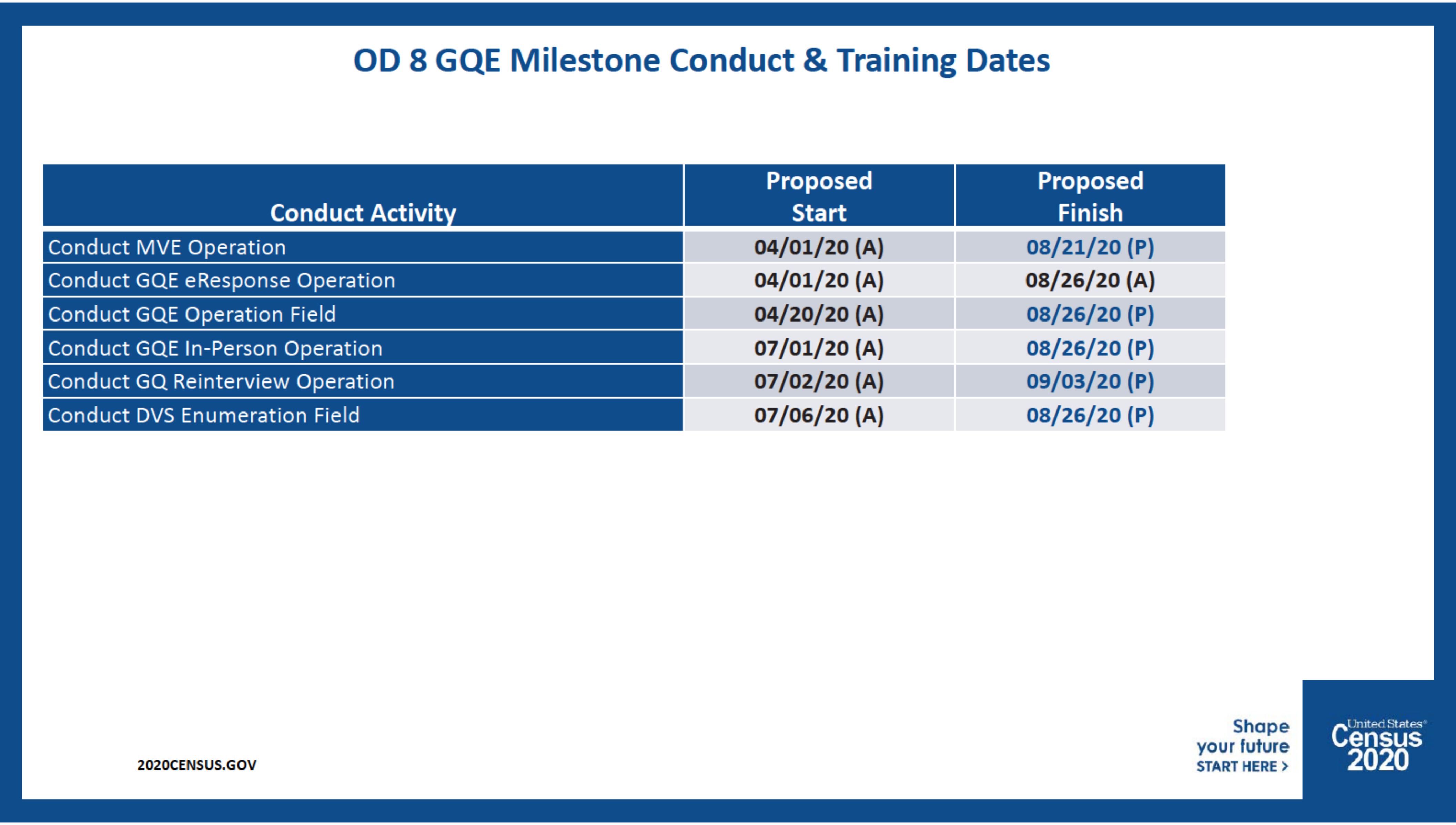


Maritime/Military Vessel Enumeration Progress

Maritime/ Military Agency	MVE Vessels in Agency	Checked In	% Vessels Checked into ATAC	# Checked-In Vessels Out of Scope	% Checked-In Vessels Out of Scope	# Vessels Enumerated	% Vessels Enumerated	# of In Progress Vessels	% of In Progress Vessels
National	1,434	1,434	100.00%	837	58.37%	1434	100.00%	0	0.00%
CFEC	645	645	100.00%	500	77.52%	645	100.00%	0	0.00%
GRNC	5	5	100.00%	1	20.00%	5	100.00%	0	0.00%
LCA	50	50	100.00%	25	50.00%	50	100.00%	0	0.00%
MARAD	235	235	100.00%	133	56.60%	235	100.00%	0	0.00%
MSC	111	111	100.00%	78	70.27%	111	100.00%	0	0.00%
NMFS	74	74	100.00%	51	68.92%	74	100.00%	0	0.00%
NOAA	15	15	100.00%	8	53.33%	15	100.00%	0	0.00%
Other Maritime/ Military	16	16	100.00%	8	50.00%	16	100.00%	0	0.00%
UNOLS	18	18	100.00%	7	38.89%	18	100.00%	0	0.00%
USCG	59	59	100.00%	5	8.47%	59	100.00%	0	0.00%
USN	206	206	100.00%	21	10.19%	206	100.00%	0	0.00%

Source: UTS Report 9/9/2020
*Per UTS, Total MVQs is 82,698

DRB Approval Number: CBDRB-FY21-DSEP-002



OD 8 GQE Milestone Conduct & Training Dates

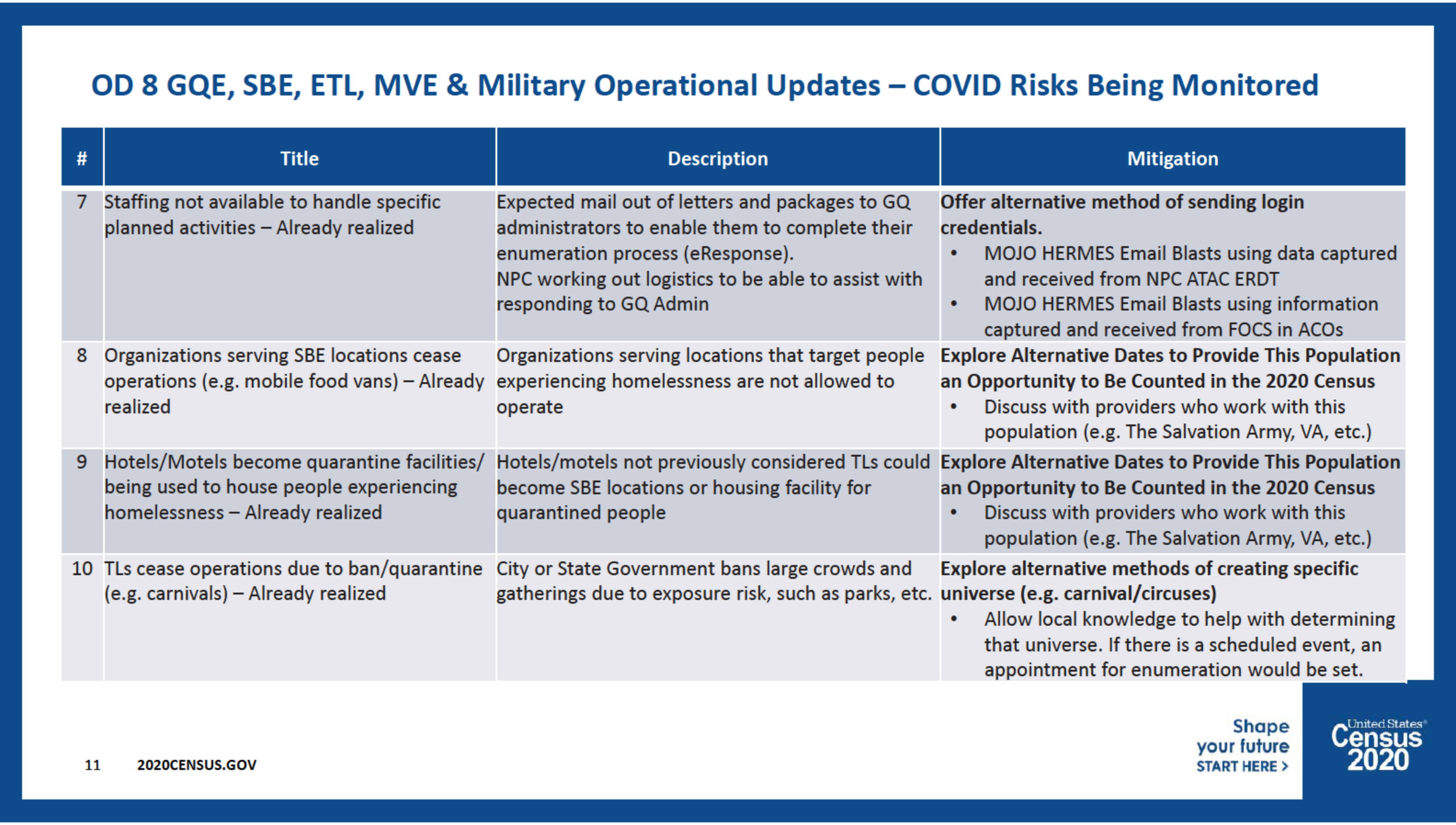
Conduct Activity	Proposed Start	Proposed Finish
Conduct MVE Operation	04/01/20 (A)	08/21/20 (P)
Conduct GQE eResponse Operation	04/01/20 (A)	08/26/20 (A)
Conduct GQE Operation Field	04/20/20 (A)	08/26/20 (P)
Conduct GQE In-Person Operation	07/01/20 (A)	08/26/20 (P)
Conduct GQ Reinterview Operation	07/02/20 (A)	09/03/20 (P)
Conduct DVS Enumeration Field	07/06/20 (A)	08/26/20 (P)

2020CENSUS.GOV

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002



OD 8 GQE, SBE, ETL, MVE & Military Operational Updates – COVID Risks Being Monitored

#	Title	Description	Mitigation
7	Staffing not available to handle specific planned activities – Already realized	Expected mail out of letters and packages to GQ administrators to enable them to complete their enumeration process (eResponse). NPC working out logistics to be able to assist with responding to GQ Admin	**Offer alternative method of sending login credentials.** • MOJO HERMES Email Blasts using data captured and received from NPC ATAC ERDT • MOJO HERMES Email Blasts using information captured and received from FOCS in ACOs
8	Organizations serving SBE locations cease operations (e.g. mobile food vans) – Already realized	Organizations serving locations that target people experiencing homelessness are not allowed to operate	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
9	Hotels/Motels become quarantine facilities/ being used to house people experiencing homelessness – Already realized	Hotels/motels not previously considered TLs could become SBE locations or housing facility for quarantined people	**Explore Alternative Dates to Provide This Population an Opportunity to Be Counted in the 2020 Census** • Discuss with providers who work with this population (e.g. The Salvation Army, VA, etc.)
10	TLs cease operations due to ban/quarantine (e.g. carnivals) – Already realized	City or State Government bans large crowds and gatherings due to exposure risk, such as parks, etc.	**Explore alternative methods of creating specific universe (e.g. carnival/circuses)** • Allow local knowledge to help with determining that universe. If there is a scheduled event, an appointment for enumeration would be set.

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

OD 8 GQE, MVE & Military Operational Updates – 2020 Census Websites

2020census.gov website

Conducting the Count: https://2020census.gov/en/conducting-the-count.html

Counting People in Group Living Arrangements: https://2020census.gov/en/conducting-the-count/gq.html

Group Quarters Enumeration: https://2020census.gov/en/conducting-the-count/gq/gqe.html

Group Quarters Advance Contact: https://2020census.gov/en/conducting-the-count/gq/gqac.html

eResponse: https://2020census.gov/en/conducting-the-count/gq/eresponse.html

Maritime and Military Vessel Enumeration: https://2020census.gov/en/conducting-the-count/gq/mve.html

Department of Education Student Privacy Policy Office: https://studentprivacy.ed.gov/faq/colleges-and-2020-census

2020 Census Operational Adjustments Due to COVID-19: https://2020census.gov/en/news-events/operational-adjustments-covid-19.html

Connect with Us

Sign up for and manage alerts at https://public.govdelivery.com/accounts/USCENSUS/subscriber/new

More information on the 2020 Census Memorandum Series: http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series.html

More information on the 2020 Census: http://www.census.gov/2020Census

More information on the American Community Survey: http://www.census.gov/programs-surveys/acs/

facebook.com/uscensusbureau

twitter.com/uscensusbureau

youtube.com/user/uscensusbureau

instagram.com/uscensusbureau

Shape your future START HERE >

Andrew Keller, Julianne Zamora, Tim Kennel
December 21, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input File:

1. [redacted].sas7bdat
2. CES 501 results
3. CES 301 results

Output File: DSSD GQ Imputation File

Section 1: Defining the Unresolved Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingesting the input File, we must initially determine what is eligible for imputation. For the cases not eligible for imputation, we assign three variables to determine this universe:
   a. **gp_initial** = This is the count of good persons in the GQ prior to imputation (0,1,….)
   b. **gpy_initial** = This indicates whether the GQ has any good persons (0/1)
   c. **unres_initial** = This indicates whether the GQ is unresolved and eligible to be imputed a positive pop count. (0/1)

```
12/21/2020
TO BEGIN: SKIP ALL the LOGIC in this Section (A) and use this:

     if GP>0 and GP PSA>0 then GP=GP PSA;
     else if GP>0 and GP PSA=. then GP=GP;
     else if GP=. and ddp in (0,.) then GP=max(CDLPER,GEO_POP_COUNT);

   if gp > 0 then gpy = 1; else gpy = 0;

 unres1 = 0;
if FOCS_ER_CB_CODE  in ('','O','R') and gpy = 0 then unres1 = 1;

unres2 = unres1;
if IMPUTE_NEEDED = 'N' then unres2 = 0;
```

```
unres=unres2;
```

1. **To determine the GQ status: start with FOCS_ER_CB_CODE**

2. **To determine the GQ has good persons (and the GQ count), I use the gp value, but I overwrite with this logic.**

```
if gp_psa > 0 then gp_initial = gp_psa
if gp_initial = . and ddp = (0,.) then gp_initial = cdlper
if gp_initial > 0 then gpy_initial = 1; else gpy_initial = 0;
```

3. **To determine the unresolved cases:**

```
unres_initial = 0;
if FOCS_ER_CB_CODE in ('','O','R ) and gpy_initial = 0 then unres_initial = 1;
```

ADK: GOTTA ADD HOW WE TAKE OUT IMPUTE_NEEDED cases and give 0 pop count if necessary

B. JEZ After making initial determinations on what is eligible for imputation, we must removed outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.
   a. **GP** = This is the count of good persons in the GQ prior to imputation (0,1,….)
   b. **GPY** = This indicates whether the GQ has any good persons (0/1)
   c. **UNRES** = This indicates whether the GQ is unresolved and eligble to be imputed an positive pop count. (0/1)

Section 2: Defining the Unresolved Cases Eligible for GQ Size Imputation

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values for when GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.
   a. We will create 3 ratios for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. If GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = ''
      i. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
      ii. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
      iii. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
      iv. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
      v. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
      vi. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
      vii. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
      viii. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
      ix. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**

Commented [JEZ(F1]: Why are these conditions on the creation of the ratios?

I would just calculate the ratios first, and then use the conditions you have to decide when to use them.

I don't understand this sub-setting. I would subset the universe for each ratio separately.

EXPRATIO = sum(GP)/sum (GQ_SIZE_EXP_PERS_CNT) where unres = '0' and FOCS_ER_CB_CODE ' ' and flagA in (' ', 'R')

MAXRATIO = sum(GP)/sum (GQ_SIZE_MAX_PERS_CNT) where unres = '0' and FOCS_ER_CB_CODE = ' ' and flagB in (' ','R')

Etc. It will be easier to code this way and it will make maximum use of the reported data.

I think you only need three sets of ratios for each of the four variables, so only 12 applicable factors for each GQTYPCUR. I think the conditions on which variables are populated only matter for the business rules at the end.

Commented [JEZ(F2R1]: I added a table at the end of the document to show what I think we should do, how we could spec out the 12 ratios.

x. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID

xi. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**

xii. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

xiii. Sum the GP and GQCURRSIZE value **for the nation.**

xiv. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

xv. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**

xvi. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID

xvii. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**

xviii. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

xix. Sum the GP and GQCURRMAXPOP value **for the nation.**

xx. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

xxi. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**

xxii. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID

xxiii. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**

xxiv. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

B. Assign Ratio-Adjustment Values for when at least one of GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT is greater than 0, but they all are not (since it is covered in the case above.

a. We will create 3 ratios for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. **<u>If it is not true that all</u>** GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0, but GQ_SIZE_EXP_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

i. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**

ii. Assign **EXPRATIO1** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

iii. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**

iv. Assign **EXPRATIO1_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID

v. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**

vi. Assign **EXPRATIO1_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

b. We will create 3 ratios for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. **<u>If it is not true that all</u>** GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0, but GQ_SIZE_MAX_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

i. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**

ii. Assign **MAXRATIO1** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

iii. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**

iv. Assign **MAXRATIO1_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID

v. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**

vi. Assign **MAXRATIO1_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

c. We will create 3 ratios for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. **If it is not true that all** GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0, but GQCURRSIZE > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

i. Sum the GP and GQCURRSIZE value for the nation.

ii. Assign **CURRSIZERATIO1** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

iii. Sum the GP and GQCURRSIZE value for each GQTYPCUR value.

iv. Assign **CURRSIZERATIO1_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID

v. Sum the GP and GQCURRSIZE value for each combination of GQTYPCUR and BCUSTATEFP value.

vi. Assign **CURRSIZERATIO1_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

d. We will create 3 ratios for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. **If it is not true that all** GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0, but GQCURRMAXPOP > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

i. Sum the GP and GQCURRMAXPOP value **for the nation.**

ii. Assign **CURRMAXRATIO1** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

iii. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**

iv. Assign **CURRMAXRATIO1_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID

v. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**

vi. Assign **CURRMAXRATIO1_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

C. Assign Good Person Percentile counts for when GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are all greater than 0.

a. We will create 3 Good Person Percentile counts for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. Do this if GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**

2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

D. Assign Good Person Percentile counts for when at least one of GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT is greater than 0, but they all are not (since it is covered in the case above.

- a. We will create 3 Good Person Percentile counts for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. Do this **<u>if it is not true that all</u>** GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0, but at least one of the four is greater than 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’
  - i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP1.**
  - ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP1_GQ.**
  - iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP1_GQ_ST.**
    1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP1_GQ_ST.**
    2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP1_GQ_ST.**
    3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP1_GQ_ST.**

E. Assign Good Person Percentile counts when GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are all 0.

- a. We will create 3 Good Person Percentile counts for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. Do this **<u>if all</u>** GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’
  - i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP0.**
  - ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP0_GQ.**
  - iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP0_GQ_ST.**
    1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP0_GQ_ST.**
    2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP0_GQ_ST.**

3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP0_GQ_ST.**

F. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

a. Define MAXPOP variable.

```
     if gqcurrmaxpop > 0 then maxpop = log(gqcurrmaxpop);
     if gqcurrmaxpop = 0 then maxpop = .;
```

b. Define the fitting universe (ratiofile) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.

d. Fit and score this model:

```
proc genmod data = ratiofile;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **poisson_count.**

G. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

H. Fold in CES 501 results

I. Fold in CES 301 results

Section 3: Applying Business Rules

The next section assigns the imputed values. It is broken into three sections based on the auxiliary data.

- GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.
- at least one of GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT is greater than 0, but they all are not (since it is covered in the case above
- GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are all 0.

A. Define these variables:

a. if GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 then hasall = 1; else hasall = 0;

b. Business Rules:

```
if GQTYPCUR = '104' and CURRSIZERATIO_GQ_ST > 0 and hasall = 1 then do;
GQIMPCT = CEIL(CURRSIZERATIO_GQ_ST * GQCURRSIZE);
GQIMPPATH = 109;
end;

else if GQTYPCUR = '104' and CURRSIZERATIO_GQ_ST <= 0 and CURRSIZERATIO_GQ > 0 and hasall
= 1 then do;
GQIMPCT = CEIL(CURRSIZERATIO_GQ * GQCURRSIZE);
GQIMPPATH = 108;
end;

else if GQTYPCUR = '104' and CURRSIZERATIO_GQ <= 0 and CURRSIZERATIO > 0 and hasall = 1
then do;
GQIMPCT = CEIL(CURRSIZERATIO * GQCURRSIZE);
GQIMPPATH = 107;
end;

if GQTYPCUR = '105' and EXPRATIO_GQ_ST > 0 and hasall = 1 then do;
GQIMPCT = CEIL(EXPRATIO_GQ_ST * GQ_SIZE_EXP_PERS_CNT);
GQIMPPATH = 103;
end;

else if GQTYPCUR = '105' and EXPRATIO_GQ_ST <= 0 and EXPRATIO_GQ > 0 and hasall = 1 then
do;
GQIMPCT = CEIL(EXPRATIO_GQ * GQ_SIZE_EXP_PERS_CNT);
GQIMPPATH = 102;
end;

else if GQTYPCUR = '105' and EXPRATIO_GQ <= 0 and EXPRATIO > 0 and hasall = 1 then do;
GQIMPCT = CEIL(EXPRATIO * GQ_SIZE_EXP_PERS_CNT);
GQIMPPATH = 101;
end;
```

GQTYPCUR	Condition (s)	Method	Flag
104	GQCURRSIZE > 0 and CURRSIZE_RATIO_GQ_ST > 0	CEIL (CURRSIZE_RATIO_GQ_ST * GQCURRSIZE)	GQIMPPATH = 1
104	GQCURRSIZE > 0 and GQCURRSIZE_GQ > 0	CEIL (CURRSIZE_RATIO_GQ * GQCURRSIZE)	GQIMPPATH = 1

Commented [JEZ(F3]: You could do a table like this and write instructions that say, do the imputation by GQTYPCUR. If a MAFID meets the set of conditions, use the method to impute the value and set the flag, if not, move to the next row, etc.

RATIOS:

Create the following ratios by summing values for all IDs where unres = 0 and FOCS_ER_CB_CODE = ‘ ‘. Use the table to determine the level for the ratio and any additional conditions. For example,

$$EXPRATIO_{GQ\_ST} = \frac{\sum_i GP}{\sum_i GQ\_SIZE\_EXP\_PERS\_CNT}$$

$$where\ i\ in\ GQTYPCUR\ and\ BCUSTATEFP\ and\ FLAGA\ in\ ('\ ','\ R')$$

Ratio	Numerator	Denominator	Level	Condition
EXPRATIO GQ ST	SUM(GP)	SUM(GQ SIZE EXP PERS CNT)	GQTYPCUR*BCUSTATEFP	FLAGA in (‘ ‘,’R’)
EXPRATIO GQ	SUM(GP)	SUM(GQ SIZE EXP PERS CNT)	GQTYPCUR	FLAGA in (‘ ‘,’R’)
EXPRATIO	SUM(GP)	SUM(GQ_SIZE_EXP_PERS_CNT)	All MAFIDs meeting conditions	FLAGA in (‘ ‘, ‘R’)
MAXRATIO GQ ST	SUM(GP)	SUM(GQ SIZE MAX PERS CNT)	GQTYPCUR*BCUSTATEFP	FLAGB in (‘ ‘,’R’)
…				

Andrew Keller, Julianne Zamora, Tim Kennel
December 23, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input Files:

1. [redacted].sas7bdat
2. [redacted].sas7bdat
3. CES 501 results
4. CES 301 results

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file, referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.

Rename GQ_INITIAL_STATUS to GQ_PRE_STATUS.
Rename GQ_INITIAL_UNRES to GQ_PRE_UNRES.
Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file, referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.
D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
    a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

Section 2: HB Edits

A. Calculate Ratios for editing.
   a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘), then
      i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT
      ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT
      iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE
      iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP
   b. Otherwise, RATIO[X] should be set to missing.
B. Create HB Parameters.
   a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR
   b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100

7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

- a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.
- b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.
- c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.
- d. For each MAFID, transform the ratio to create **SVALUE**.
    - i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])
    - ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)
- e. For each MAFID, transform SVALUE to create **EVALUE**.
    - i. EVALUE = SVALUE * max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}$^{0.5}$
    - ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.
- f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.
    - i. **E_Q1** = first quartile EVALUE
    - ii. **E_MED** = median EVALUE
    - iii. **E_Q3** = third quartile EVALUE
- g. For each GQTYPE, define upper and lower bounds.
    - i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}
    - ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}
    - iii. **LOWER_C1** = E_MED – C1 * D_Q1
    - iv. **LOWER_C2** = E_MED – C2 * D_Q1
    - v. **LOWER_C3** = E_MED – C3 * D_Q1
    - vi. **UPPER_C1** = E_MED + C1 * D_Q3
    - vii. **UPPER_C2** = E_MED + C2 * D_Q3
    - viii. **UPPER_C3** = E_MED + C3 * D_Q3
- h. For each MAFID, create **FLAG[X]**.
    - i. If EVALUE is missing, FLAG[X] = 'M'
    - ii. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = 'R'
    - iii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'
    - iv. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

- a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0. Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:
   i. If FLAG[X] = ‘I’ and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = ‘S’.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.
   a. If (FLAGA = ‘I’ or FLAGB = ‘I’ or FLAGC = ‘I’ or FLAGD = ‘I’) and IMPUTE_NEEDED ne ‘N’ then
      i. **GP = .**
      ii. **UNRES** = 1
   b. Otherwise,
      i. **GP =** GQ_INITIAL_POP
      ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values
   a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values
      i. Calculate Ratios.
         We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’):
         1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
         2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
         3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
         4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
         5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
         6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
      ii. Assign values. For each MAFID, calculate the following values:
         1. **IMP_RAT_EXP** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO)
         2. **IMP_RAT_EXP_GQ** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ)
         3. **IMP_RAT_EXP_GQ_ST** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values
    i. Calculate Ratios.
    We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):
        1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
        2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
        3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
        4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
        5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
        6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
    ii. Assign values. For each MAFID, calculate the following values:
        1. **IMP_RAT_MAX** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO)
        2. **IMP_RAT_MAX_GQ** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ)
        3. **IMPRAT_MAX_GQ_ST** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values
    i. Calculate Ratios.
    We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):
        1. Sum the GP and GQCURRSIZE value **for the nation.**
        2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
        3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
        4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
        5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
        6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
    ii. Assign values. For each MAFID, calculate the following values:
        1. **IMP_RAT_CURR** = CEIL (GQCURRSIZE*CURRSIZERATIO)
        2. **IMP_RAT_CURR_GQ** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ)
        3. **IMP_RAT_CURR_GQ_ST** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values
    i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (GQCURRMAXPOP*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **<u>all</u>** greater than 0.

a. Define MAXPOP variable.

i. if GQCURRMAXPOP > 0 then **MAXPOP** = log(GQCURRMAXPOP);

ii. if GQCURRMAXPOP = 0 then **MAXPOP** = .;

b. Define the fitting universe (ratiofile) as this: FLAGA in (' ','R') and FLAGB in (' ','R') and FLAGC in (' ','R') and FLAGD in (' ','R') and unres = 0 and FOCS ER CB CODE = ''

c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.

d. Fit and score this model:

```
proc genmod data = ratiofile;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **IMP_POISSON_COUNT.**

Commented [JEZ(F1]: Remove?

D. Fold in CES 501 results

Commented [JEZ(F2]: Residual Method

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table as follows, if IMP_POISSON_COUNT is not missing, assign IMP_GP = IMP_POISSON_COUNT and assign IMP_FLAG = 201. If IMP_POISSON_COUNT is missing, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. Continue on through the table until all MAFIDs in unres = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	IMP_FLAG
IMP POISSON COUNT	201
IMP RAT EXP GQ ST	101
IMP RAT EXP GQ	102
IMP RAT EXP	103
IMP RAT MAX GQ ST	104
IMP RAT MAX GQ	105
IMP RAT MAX	106
IMP RAT CURR GQ ST	107
IMP RAT CURR GQ	108
'IMP RAT CURR	109
IMP RAT CURRMAX GQ ST	110
IMP RAT CURRMAX GQ	111
IMP RAT CURRMAX	112
MEDGP GQ ST	401
MEDGP GQ	402
MEDGP	403

Commented [JEZ(F3]: Remove?

<u>Section 6: Create Output File</u>

Output GQ MAFID, adding the following variables:

FLAGA	FLAGB	
FLAGC	FLAGD	
GP	UNRES	
EXPRATIO	EXPRATIO GQ	EXPRATIO GQ ST
IMP RAT EXP	IMP RAT EXP GQ	IMP RAT EXP GQ ST
MAXRATIO	MAXRATIO GQ	MAXRATIO GQ ST
IMP RAT MAX	IMP RAT MAX GQ	IMP RAT MAX GQ ST
CURRRATIO	CURRRATIO GQ	CURRATIO GQ ST
IMP RAT CURR	IMP RAT CURR GQ	IMP RAT CURR GQ ST
MAXCURRRATIO	MAXCURRRATIO GQ	MAXCURRRATIO GQ ST
IMP RAT MAXCURR	IMP RAT MAXCURR GQ	IMP RAT MAXCURR GQ ST
MEDGP	MEDGP GQ	MEDGP GQ ST
IMP GP	IMP FLAG	

Name this file gq_mafid_dssd_out.sas7bdat

Andrew Keller, Julianne Zamora, Tim Kennel
December 21, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input File:

1. [redacted].sas7bdat
2. CES 501 results
3. CES 301 results

Output File: DSSD GQ Imputation File

Section 1: Defining the Unresolved Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingesting the input File, we must initially determine what is eligible for imputation. For the cases not eligible for imputation, we assign three variables to determine this universe:
    a. **gp_initial** = This is the count of good persons in the GQ prior to imputation (0,1,….)
    b. **gpy_initial** = This indicates whether the GQ has any good persons (0/1)
    c. **unres_initial** = This indicates whether the GQ is unresolved and eligible to be imputed a positive pop count. (0/1)

```
12/21/2020
TO BEGIN: SKIP ALL the LOGIC in this Section (A) and use this:

     if GP>0 and GP PSA>0 then GP=GP PSA;
     else if GP>0 and GP PSA=. then GP=GP;
     else if GP=. and ddp in (0,.) then GP=max(CDLPER,GEO_POP_COUNT);

   if gp > 0 then gpy = 1; else gpy = 0;

 unres1 = 0;
if FOCS_ER_CB_CODE  in ('','O','R') and gpy = 0 then unres1 = 1;

unres2 = unres1;
if IMPUTE_NEEDED = 'N' then unres2 = 0;
```

```
unres=unres2;
```

1. **To determine the GQ status: start with FOCS_ER_CB_CODE**

2. **To determine the GQ has good persons (and the GQ count), I use the gp value, but I overwrite with this logic.**

   if gp_psa > 0 then gp_initial = gp_psa

   if gp_initial = . and ddp = (0,.) then gp_initial = cdlper

   if gp_initial > 0 then gpy_initial = 1; else gpy_initial = 0;

3. **To determine the unresolved cases:**

   unres_initial = 0;

   if FOCS_ER_CB_CODE in ('','O','R) and gpy_initial = 0 then unres_initial = 1;

   ADK: GOTTA ADD HOW WE TAKE OUT IMPUTE_NEEDED cases and give 0 pop count if necessary

B. JEZ After making initial determinations on what is eligible for imputation, we must removed outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

   a. **GP** = This is the count of good persons in the GQ prior to imputation (0,1,....)
   b. **GPY** = This indicates whether the GQ has any good persons (0/1)
   c. **UNRES** = This indicates whether the GQ is unresolved and eligble to be imputed an positive pop count. (0/1)

Section 2: Defining the Unresolved Cases Eligible for GQ Size Imputation

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

   a. We will create 3 ratios for each variable, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):
      i. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
      ii. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
      iii. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
      iv. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
      v. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
      vi. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
   b. We will create 3 ratios for each variable, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):

Commented [JEZ(F1]: Why are these conditions on the creation of the ratios?

I would just calculate the ratios first, and then use the conditions you have to decide when to use them.

I don't understand this sub-setting. I would subset the universe for each ratio separately.

EXPRATIO = sum(GP)/sum (GQ_SIZE_EXP_PERS_CNT) where unres = '0' and FOCS_ER_CB_CODE ' ' and flagA in (' ', 'R')

MAXRATIO = sum(GP)/sum (GQ_SIZE_MAX_PERS_CNT) where unres = '0' and FOCS_ER_CB_CODE = ' ' and flagB in (' ','R')

Etc. It will be easier to code this way and it will make maximum use of the reported data.

I think you only need three sets of ratios for each of the four variables, so only 12 applicable factors for each GQTYPCUR. I think the conditions on which variables are populated only matter for the business rules at the end.

Commented [JEZ(F2R1]: I added a table at the end of the document to show what I think we should do, how we could spec out the 12 ratios.

i. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
ii. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
iii. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
iv. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
v. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
vi. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

c. We will create 3 ratios for each variable, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagC in (‘’,’R’):
   i. Sum the GP and GQCURRSIZE value **for the nation.**
   ii. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
   iii. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
   iv. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
   v. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
   vi. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

d. We will create 3 ratios for each variable, one for the national value (**CURRMAXRATIO)**, one for the GQTYPCUR combination (**CURRMAXRATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagD in (‘’,’R’):
   i. Sum the GP and GQCURRMAXPOP value **for the nation.**
   ii. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
   iii. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
   iv. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
   v. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
   vi. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

B. Assign Good Person Percentile counts.
   a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP)**, one for the GQTYPCUR combination (**MEDGP_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’,’R’):
      i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**
      ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**
      iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**
         1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**

2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **<u>all</u>** greater than 0.

a. Define MAXPOP variable.

```
      if gqcurrmaxpop > 0 then maxpop = log(gqcurrmaxpop);
      if gqcurrmaxpop = 0 then maxpop = .;
```

b. Define the fitting universe (ratiofile) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = “

c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.

d. Fit and score this model:

```
proc genmod data = ratiofile;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **poisson_count.**

D. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **<u>all</u>** greater than 0.

E. Fold in CES 501 results

F. Fold in CES 301 results

<u>Section 3: Applying Business Rules</u>

The next section assigns the imputed values. It is broken into three sections based on the auxiliary data.

A. GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **<u>all</u>** greater than 0.

GQTYPCUR	**Condition (s)**	**Method**	**Flag**

Commented [JEZ(F3]: You could do a table like this and write instructions that say, do the imputation by GQTYPCUR. If a MAFID meets the set of conditions, use the method to impute the value and set the flag, if not, move to the next row, etc.

104	GQCURRSIZE > 0 and CURRSIZE_RATIO_GQ_ST > 0	CEIL (CURRSIZE_RATIO_GQ_ST * GQCURRSIZE)	GQIMPPATH = 109
104	GQCURRSIZE > 0 and GQCURRSIZE_GQ > 0	CEIL (CURRSIZE_RATIO_GQ * GQCURRSIZE)	GQIMPPATH = 108

B. at least one of GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT is greater than 0, but they all are not (since it is covered in the case above

C. GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are all 0.

RATIOS:

Create the following ratios by summing values for all IDs where unres = 0 and FOCS_ER_CB_CODE = ‘ ‘. Use the table to determine the level for the ratio and any additional conditions. For example,

$$EXPRATIO_{GQ\_ST} = \frac{\sum_i GP}{\sum_i GQ\_SIZE\_EXP\_PERS\_CNT}$$

$$where\ i\ in\ GQTYPCUR\ and\ BCUSTATEFP\ and\ FLAGA\ in\ ('\ ','\ R')$$

Ratio	Numerator	Denominator	Level	Condition
EXPRATIO GQ ST	SUM(GP)	SUM(GQ SIZE EXP PERS CNT)	GQTYPCUR*BCUSTATEFP	FLAGA in (‘ ‘,’R’)
EXPRATIO GQ	SUM(GP)	SUM(GQ SIZE EXP PERS CNT)	GQTYPCUR	FLAGA in (‘ ‘,’R’)
EXPRATIO	SUM(GP)	SUM(GQ_SIZE_EXP_PERS_CNT)	All MAFIDs meeting conditions	FLAGA in (‘ ‘, ‘R’)
MAXRATIO GQ ST	SUM(GP)	SUM(GQ SIZE MAX PERS CNT)	GQTYPCUR*BCUSTATEFP	FLAGB in (‘ ‘,’R’)
…				

Andrew Keller, Julianne Zamora, Tim Kennel
December 21, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input File:

1. [redacted].sas7bdat
2. CES 501 results
3. CES 301 results

Output File: DSSD GQ Imputation File

Section 1: Defining the Unresolved Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingesting the input File, we must initially determine what is eligible for imputation. For the cases not eligible for imputation, we assign three variables to determine this universe:
   a. **gp_initial** = This is the count of good persons in the GQ prior to imputation (0,1,….)
   b. **gpy_initial** = This indicates whether the GQ has any good persons (0/1)
   c. **unres_initial** = This indicates whether the GQ is unresolved and eligible to be imputed a positive pop count. (0/1)

```
12/21/2020
TO BEGIN: SKIP ALL the LOGIC in this Section (A) and use this:

     if GP>0 and GP PSA>0 then GP=GP PSA;
     else if GP>0 and GP PSA=. then GP=GP;
     else if GP=. and ddp in (0,.) then GP=max(CDLPER,GEO_POP_COUNT);

   if gp > 0 then gpy = 1; else gpy = 0;

 unres1 = 0;
if FOCS_ER_CB_CODE  in ('','O','R') and gpy = 0 then unres1 = 1;

unres2 = unres1;
if IMPUTE_NEEDED = 'N' then unres2 = 0;
```

```
unres=unres2;
```

1. **To determine the GQ status: start with FOCS_ER_CB_CODE**

2. **To determine the GQ has good persons (and the GQ count), I use the gp value, but I overwrite with this logic.**

   if gp_psa > 0 then gp_initial = gp_psa
   if gp_initial = . and ddp = (0,.) then gp_initial = cdlper
   if gp_initial > 0 then gpy_initial = 1; else gpy_initial = 0;

3. **To determine the unresolved cases:**

   unres_initial = 0;
   if FOCS_ER_CB_CODE in ('','O','R) and gpy_initial = 0 then unres_initial = 1;
   ADK: GOTTA ADD HOW WE TAKE OUT IMPUTE_NEEDED cases and give 0 pop count if necessary

B. JEZ After making initial determinations on what is eligible for imputation, we must removed outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

   a. **GP** = This is the count of good persons in the GQ prior to imputation (0,1,....)
   b. **GPY** = This indicates whether the GQ has any good persons (0/1)
   c. **UNRES** = This indicates whether the GQ is unresolved and eligble to be imputed an positive pop count. (0/1)

Section 2: Defining the Unresolved Cases Eligible for GQ Size Imputation

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

   a. We will create 3 ratios for each variable, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):
      i. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
      ii. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
      iii. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
      iv. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
      v. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
      vi. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

   b. We will create 3 ratios for each variable, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):

Commented [JEZ(F1]: Why are these conditions on the creation of the ratios?

I would just calculate the ratios first, and then use the conditions you have to decide when to use them.

I don't understand this sub-setting. I would subset the universe for each ratio separately.

EXPRATIO = sum(GP)/sum (GQ_SIZE_EXP_PERS_CNT) where unres = '0' and FOCS_ER_CB_CODE ' ' and flagA in (' ', 'R')

MAXRATIO = sum(GP)/sum (GQ_SIZE_MAX_PERS_CNT) where unres = '0' and FOCS_ER_CB_CODE = ' ' and flagB in (' ','R')

Etc. It will be easier to code this way and it will make maximum use of the reported data.

I think you only need three sets of ratios for each of the four variables, so only 12 applicable factors for each GQTYPCUR. I think the conditions on which variables are populated only matter for the business rules at the end.

Commented [JEZ(F2R1]: I added a table at the end of the document to show what I think we should do, how we could spec out the 12 ratios.

i. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
ii. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
iii. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
iv. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
v. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
vi. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

c. We will create 3 ratios for each variable, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):
i. Sum the GP and GQCURRSIZE value **for the nation.**
ii. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
iii. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
iv. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
v. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
vi. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

d. We will create 3 ratios for each variable, one for the national value (**CURRMAXRATIO)**, one for the GQTYPCUR combination (**CURRMAXRATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):
i. Sum the GP and GQCURRMAXPOP value **for the nation.**
ii. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
iii. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
iv. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
v. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
vi. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

B. Assign Good Person Percentile counts.
a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP)**, one for the GQTYPCUR combination (**MEDGP_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):
i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**
ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**
iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**
1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**

2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

a. Define MAXPOP variable.

```
      if gqcurrmaxpop > 0 then maxpop = log(gqcurrmaxpop);
      if gqcurrmaxpop = 0 then maxpop = .;
```

b. Define the fitting universe (ratiofile) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = "

c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.

d. Fit and score this model:

```
proc genmod data = ratiofile;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **poisson_count.**

D. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

E. Fold in CES 501 results

F. Fold in CES 301 results

Section 3: Applying Business Rules

The next section assigns the imputed values. It is broken into three sections based on the auxiliary data.

A. GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

GQTYPCUR	Condition (s)	Method	Flag

Commented [JEZ(F3]: You could do a table like this and write instructions that say, do the imputation by GQTYPCUR. If a MAFID meets the set of conditions, use the method to impute the value and set the flag, if not, move to the next row, etc.

104	GQCURRSIZE > 0 and CURRSIZE_RATIO_GQ_ST > 0	CEIL (CURRSIZE_RATIO_GQ_ST * GQCURRSIZE)	GQIMPPATH = 109
104	GQCURRSIZE > 0 and GQCURRSIZE_GQ > 0	CEIL (CURRSIZE_RATIO_GQ * GQCURRSIZE)	GQIMPPATH = 108

B. at least one of GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT is greater than 0, but they all are not (since it is covered in the case above

C. GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are all 0.

RATIOS:

Create the following ratios by summing values for all IDs where unres = 0 and FOCS_ER_CB_CODE = ‘ ‘. Use the table to determine the level for the ratio and any additional conditions. For example,

$$EXPRATIO_{GQ\_ST} = \frac{\sum_i GP}{\sum_i GQ\_SIZE\_EXP\_PERS\_CNT}$$

$$where\ i\ in\ GQTYPCUR\ and\ BCUSTATEFP\ and\ FLAGA\ in\ ('\ ','\ R')$$

Ratio	Numerator	Denominator	Level	Condition
EXPRATIO GQ ST	SUM(GP)	SUM(GQ SIZE EXP PERS CNT)	GQTYPCUR*BCUSTATEFP	FLAGA in (‘ ‘,’R’)
EXPRATIO GQ	SUM(GP)	SUM(GQ SIZE EXP PERS CNT)	GQTYPCUR	FLAGA in (‘ ‘,’R’)
EXPRATIO	SUM(GP)	SUM(GQ_SIZE_EXP_PERS_CNT)	All MAFIDs meeting conditions	FLAGA in (‘ ‘, ‘R’)
MAXRATIO GQ ST	SUM(GP)	SUM(GQ SIZE MAX PERS CNT)	GQTYPCUR*BCUSTATEFP	FLAGB in (‘ ‘,’R’)
…				

Andrew Keller, Julianne Zamora, Tim Kennel
December 23, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input Files:

1. [redacted].sas7bdat
2. [redacted].sas7bdat
3. CES 501 results
4. CES 301 results

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file, referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.

Rename GQ_INITIAL_STATUS to GQ_PRE_STATUS.
Rename GQ_INITIAL_UNRES to GQ_PRE_UNRES.
Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file, referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.
D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
    a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

Section 2: HB Edits

A. Calculate Ratios for editing.

a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘), then

i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT

ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT

iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE

iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP

b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.

a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR

b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100

7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

  a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

  b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

  c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

  d. For each MAFID, transform the ratio to create **SVALUE**.

    i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

    ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)

  e. For each MAFID, transform SVALUE to create **EVALUE**.

    i. EVALUE = SVALUE * max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}$^{0.5}$

    ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

  f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

    i. **E_Q1** = first quartile EVALUE

    ii. **E_MED** = median EVALUE

    iii. **E_Q3** = third quartile EVALUE

  g. For each GQTYPE, define upper and lower bounds.

    i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

    ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

    iii. **LOWER_C1** = E_MED – C1 * D_Q1

    iv. **LOWER_C2** = E_MED – C2 * D_Q1

    v. **LOWER_C3** = E_MED – C3 * D_Q1

    vi. **UPPER_C1** = E_MED + C1 * D_Q3

    vii. **UPPER_C2** = E_MED + C2 * D_Q3

    viii. **UPPER_C3** = E_MED + C3 * D_Q3

  h. For each MAFID, create **FLAG[X]**.

    i. If EVALUE is missing, FLAG[X] = 'M'

    ii. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = 'R'

    iii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'

    iv. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

  a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0. Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:
   i. If FLAG[X] = ‘I’ and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = ‘S’.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.
   a. If (FLAGA = ‘I’ or FLAGB = ‘I’ or FLAGC = ‘I’ or FLAGD = ‘I’) and IMPUTE_NEEDED ne ‘N’ then
      i. **GP = .**
      ii. **UNRES** = 1
   b. Otherwise,
      i. **GP =** GQ_INITIAL_POP
      ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values
   a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values
      i. Calculate Ratios.
         We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’):
         1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
         2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
         3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
         4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
         5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
         6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
      ii. Assign values. For each MAFID, calculate the following values:
         1. **IMP_RAT_EXP** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO)
         2. **IMP_RAT_EXP_GQ** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ)
         3. **IMP_RAT_EXP_GQ_ST** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values

   i. Calculate Ratios.
      We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):

      1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
      2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
      3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
      4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
      5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
      6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

   ii. Assign values. For each MAFID, calculate the following values:

      1. **IMP_RAT_MAX** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO)
      2. **IMP_RAT_MAX_GQ** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ)
      3. **IMPRAT_MAX_GQ_ST** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

   i. Calculate Ratios.
      We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):

      1. Sum the GP and GQCURRSIZE value **for the nation.**
      2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
      3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
      4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
      5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
      6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

   ii. Assign values. For each MAFID, calculate the following values:

      1. **IMP_RAT_CURR** = CEIL (GQCURRSIZE*CURRSIZERATIO)
      2. **IMP_RAT_CURR_GQ** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ)
      3. **IMP_RAT_CURR_GQ_ST** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

   i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (GQCURRMAXPOP*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **<u>all</u>** greater than 0.

a. Define MAXPOP variable.

i. if GQCURRMAXPOP > 0 then **MAXPOP** = log(GQCURRMAXPOP);

ii. if GQCURRMAXPOP = 0 then **MAXPOP** = .;

b. Define the fitting universe (ratiofile) as this: FLAGA in (' ','R') and FLAGB in (' ','R') and FLAGC in (' ','R') and FLAGD in (' ','R') and unres = 0 and FOCS ER CB CODE = ''

c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.

d. Fit and score this model:

```
proc genmod data = ratiofile;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **IMP_POISSON_COUNT.**

Commented [JEZ(F1]: Remove?

D. Fold in CES 501 results

Commented [JEZ(F2]: Residual Method

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table as follows, if IMP_POISSON_COUNT is not missing, assign IMP_GP = IMP_POISSON_COUNT and assign IMP_FLAG = 201. If IMP_POISSON_COUNT is missing, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. Continue on through the table until all MAFIDs in unres = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	IMP_FLAG
IMP POISSON COUNT	201
IMP RAT EXP GQ ST	101
IMP RAT EXP GQ	102
IMP RAT EXP	103
IMP RAT MAX GQ ST	104
IMP RAT MAX GQ	105
IMP RAT MAX	106
IMP RAT CURR GQ ST	107
IMP RAT CURR GQ	108
'IMP RAT CURR	109
IMP RAT CURRMAX GQ ST	110
IMP RAT CURRMAX GQ	111
IMP RAT CURRMAX	112
MEDGP GQ ST	401
MEDGP GQ	402
MEDGP	403

Commented [JEZ(F3]: Remove?

Section 6: Create Output File

Output GQ MAFID, adding the following variables:

FLAGA	FLAGB	
FLAGC	FLAGD	
GP	UNRES	
EXPRATIO	EXPRATIO GQ	EXPRATIO GQ ST
IMP RAT EXP	IMP RAT EXP GQ	IMP RAT EXP GQ ST
MAXRATIO	MAXRATIO GQ	MAXRATIO GQ ST
IMP RAT MAX	IMP RAT MAX GQ	IMP RAT MAX GQ ST
CURRRATIO	CURRRATIO GQ	CURRATIO GQ ST
IMP RAT CURR	IMP RAT CURR GQ	IMP RAT CURR GQ ST
MAXCURRRATIO	MAXCURRRATIO GQ	MAXCURRRATIO GQ ST
IMP RAT MAXCURR	IMP RAT MAXCURR GQ	IMP RAT MAXCURR GQ ST
MEDGP	MEDGP GQ	MEDGP GQ ST
IMP GP	IMP FLAG	

Name this file gq_mafid_dssd_out.sas7bdat

Andrew Keller, Julianne Zamora, Tim Kennel
December 23, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input Files:

1. [redacted].sas7bdat
2. [redacted].sas7bdat
3. CES 501 results
4. CES 301 results

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file, referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.

Section 2: HB Edits

A. Calculate Ratios for editing.
    a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in ('O','R',' ') AND GQ_INITIAL_POP > 0 then
        i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT
        ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT
        iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE
        iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP
    b. Otherwise, RATIO[X] should be set to missing.
B. Create HB Parameters.

a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR

b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100
7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

d. For each MAFID, transform the ratio to create **SVALUE**.

  i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

  ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)

e. For each MAFID, transform SVALUE to create **EVALUE**.

  i. EVALUE = SVALUE * max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}$^{0.5}$

  ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

  i. **E_Q1** = first quartile EVALUE

  ii. **E_MED** = median EVALUE

  iii. **E_Q3** = third quartile EVALUE

g. For each GQTYPE, define upper and lower bounds.

  i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

  ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

  iii. **LOWER_C1** = E_MED – C1 * D_Q1

  iv. **LOWER_C2** = E_MED – C2 * D_Q1

  v. **LOWER_C3** = E_MED – C3 * D_Q1

  vi. **UPPER_C1** = E_MED + C1 * D_Q3

  vii. **UPPER_C2** = E_MED + C2 * D_Q3

  viii. **UPPER_C3** = E_MED + C3 * D_Q3

h. For each MAFID, create **FLAG[X]**.

  i. If EVALUE is missing, FLAG[X] = 'M'

  ii. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = 'R'

  iii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'

  iv. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0 AND FLAGA not in ('S','I') and FLAGB not in ('S','I') and FLAGC not in ('S','I') and FLAGD not in ('S','I'). Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

  i. If FLAG[X] = 'I' and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = 'S'.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

  a. If (FLAGA = 'I' or FLAGB = 'I' or FLAGC = 'I' or FLAGD = 'I') and IMPUTE_NEEDED ne 'N' then
     i. **GP = .**
     ii. **UNRES** = 1
  b. Otherwise,
     i. **GP =** GQ_INITIAL_POP
     ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values
  a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values
     i. Calculate Ratios.
        We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):
        1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
        2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
        3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
        4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
        5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
        6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
     ii. Assign values. For each MAFID, calculate the following values:
        1. **IMP_RAT_EXP** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO)
        2. **IMP_RAT_EXP_GQ** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ)
        3. **IMP_RAT_EXP_GQ_ST** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ_ST)

  b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values
     i. Calculate Ratios.
        We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):
        1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**

2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_MAX** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO)
2. **IMP_RAT_MAX_GQ** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ)
3. **IMPRAT_MAX_GQ_ST** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagC in (‘’,’R’):

1. Sum the GP and GQCURRSIZE value **for the nation.**
2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURR** = CEIL (GQCURRSIZE*CURRSIZERATIO)
2. **IMP_RAT_CURR_GQ** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ)
3. **IMP_RAT_CURR_GQ_ST** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagD in (‘’,’R’):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**

4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (GQCURRMAXPOP*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’,’R’):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are <u>**all**</u> greater than 0.

a. Define MAXPOP variable.

i. if GQCURRMAXPOP > 0 then **MAXPOP** = log(GQCURRMAXPOP);

ii. if GQCURRMAXPOP = 0 then **MAXPOP** = .;

b. Define the fitting universe (ratiofile) as this: FLAGA in (‘ ‘,’R’) and FLAGB in (‘ ‘,’R’) and FLAGC in (‘ ‘,’R’) and FLAGD in (‘ ‘,’R’) and unres = 0 and FOCS ER CB CODE = ‘’

c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.

d. Fit and score this model:

```
proc genmod data = ratiofile;
    class gqtypcur;
    model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
```

```
          link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **IMP_POISSON_COUNT.**

Commented [JEZ(F1]: Remove?

D. Fold in CES 501 results

Commented [JEZ(F2]: Residual Method

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table as follows, if IMP_POISSON_COUNT is not missing, assign IMP_GP = IMP_POISSON_COUNT and assign IMP_FLAG = 201. If IMP_POISSON_COUNT is missing, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. Continue on through the table until all MAFIDs in unres = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	IMP_FLAG
IMP POISSON COUNT	201
IMP RAT EXP GQ ST	101
IMP RAT EXP GQ	102
IMP RAT EXP	103
IMP RAT MAX GQ ST	104
IMP RAT MAX GQ	105
IMP RAT MAX	106
IMP RAT CURR GQ ST	107
IMP RAT CURR GQ	108
IMP RAT CURR	109
IMP RAT MAXCURR GQ ST	110
IMP RAT MAXCURR GQ	111
IMP RAT MAXCURR	112
MEDGP GQ ST	401
MEDGP GQ	402
MEDGP	403

Commented [JEZ(F3]: Remove?

Section 6: Create Output File

Output GQ MAFID, adding the following variables, renaming GP to GP HB:

FLAGA	FLAGB	
FLAGC	FLAGD	
GP HB	UNRES	
EXPRATIO	EXPRATIO GQ	EXPRATIO GQ ST
IMP RAT EXP	IMP RAT EXP GQ	IMP RAT EXP GQ ST

MAXRATIO	MAXRATIO GQ	MAXRATIO GQ ST
IMP RAT MAX	IMP RAT MAX GQ	IMP RAT MAX GQ ST
CURRRATIO	CURRRATIO GQ	CURRATIO GQ ST
IMP RAT CURR	IMP RAT CURR GQ	IMP RAT CURR GQ ST
MAXCURRRATIO	MAXCURRRATIO GQ	MAXCURRRATIO GQ ST
IMP RAT MAXCURR	IMP RAT MAXCURR GQ	IMP RAT MAXCURR GQ ST
MEDGP	MEDGP GQ	MEDGP GQ ST
IMP GP	IMP FLAG	

Name this file gq_mafid_dssd_out.sas7bdat

Andrew Keller, Julianne Zamora, Tim Kennel
December 23, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input Files:

1. [redacted].sas7bdat
2. [redacted].sas7bdat
3. CES 501 results
4. CES 301 results

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file, referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.

Rename GQ_INITIAL_STATUS to GQ_PRE_STATUS.
Rename GQ_INITIAL_UNRES to GQ_PRE_UNRES.
Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file, referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.
D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
    a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

<u>Section 2: HB Edits</u>

A. Calculate Ratios for editing.

   a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘), then

      i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT

      ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT

      iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE

      iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP

   b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.

   a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR

   b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100

7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

- a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.
- b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.
- c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.
- d. For each MAFID, transform the ratio to create **SVALUE**.
    - i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])
    - ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)
- e. For each MAFID, transform SVALUE to create **EVALUE**.
    - i. EVALUE = SVALUE * max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}$^{0.5}$
    - ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.
- f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.
    - i. **E_Q1** = first quartile EVALUE
    - ii. **E_MED** = median EVALUE
    - iii. **E_Q3** = third quartile EVALUE
- g. For each GQTYPE, define upper and lower bounds.
    - i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}
    - ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}
    - iii. **LOWER_C1** = E_MED – C1 * D_Q1
    - iv. **LOWER_C2** = E_MED – C2 * D_Q1
    - v. **LOWER_C3** = E_MED – C3 * D_Q1
    - vi. **UPPER_C1** = E_MED + C1 * D_Q3
    - vii. **UPPER_C2** = E_MED + C2 * D_Q3
    - viii. **UPPER_C3** = E_MED + C3 * D_Q3
- h. For each MAFID, create **FLAG[X]**.
    - i. If EVALUE is missing, FLAG[X] = 'M'
    - ii. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = 'R'
    - iii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'
    - iv. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

- a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0. Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

i. If FLAG[X] = ‘I’ and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = ‘S’.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

a. If (FLAGA = ‘I’ or FLAGB = ‘I’ or FLAGC = ‘I’ or FLAGD = ‘I’) and IMPUTE_NEEDED ne ‘N’ then

i. **GP = .**

ii. **UNRES** = 1

b. Otherwise,

i. **GP =** GQ_INITIAL_POP

ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’):

1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_EXP** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO)
2. **IMP_RAT_EXP_GQ** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ)
3. **IMP_RAT_EXP_GQ_ST** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values
    i. Calculate Ratios.
    We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):
        1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
        2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
        3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
        4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
        5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
        6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
    ii. Assign values. For each MAFID, calculate the following values:
        1. **IMP_RAT_MAX** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO)
        2. **IMP_RAT_MAX_GQ** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ)
        3. **IMPRAT_MAX_GQ_ST** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values
    i. Calculate Ratios.
    We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):
        1. Sum the GP and GQCURRSIZE value **for the nation.**
        2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
        3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
        4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
        5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
        6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
    ii. Assign values. For each MAFID, calculate the following values:
        1. **IMP_RAT_CURR** = CEIL (GQCURRSIZE*CURRSIZERATIO)
        2. **IMP_RAT_CURR_GQ** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ)
        3. **IMP_RAT_CURR_GQ_ST** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values
    i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (GQCURRMAXPOP*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **<u>all</u>** greater than 0.

a. Define MAXPOP variable.

i. if GQCURRMAXPOP > 0 then **MAXPOP** = log(GQCURRMAXPOP);

ii. if GQCURRMAXPOP = 0 then **MAXPOP** = .;

b. Define the fitting universe (ratiofile) as this: FLAGA in (' ','R') and FLAGB in (' ','R') and FLAGC in (' ','R') and FLAGD in (' ','R') and unres = 0 and FOCS ER CB CODE = ''

c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.

d. Fit and score this model:

```
proc genmod data = ratiofile;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **IMP_POISSON_COUNT.**

Commented [JEZ(F1]: Remove?

D. Fold in CES 501 results

Commented [JEZ(F2]: Residual Method

<u>Section 5: Apply Ordering to Select Final Imputed Value</u>

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table as follows, if IMP_POISSON_COUNT is not missing, assign IMP_GP = IMP_POISSON_COUNT and assign IMP_FLAG = 201. If IMP_POISSON_COUNT is missing, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. Continue on through the table until all MAFIDs in unres = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	IMP_FLAG
IMP POISSON COUNT	201
IMP RAT EXP GQ ST	101
IMP RAT EXP GQ	102
IMP RAT EXP	103
IMP RAT MAX GQ ST	104
IMP RAT MAX GQ	105
IMP RAT MAX	106
IMP RAT CURR GQ ST	107
IMP RAT CURR GQ	108
'IMP RAT CURR	109
IMP RAT CURRMAX GQ ST	110
IMP RAT CURRMAX GQ	111
IMP RAT CURRMAX	112
MEDGP GQ ST	401
MEDGP GQ	402
MEDGP	403

Commented [JEZ(F3]: Remove?

Section 6: Create Output File

Output GQ MAFID, adding the following variables:

FLAGA	FLAGB	
FLAGC	FLAGD	
GP	UNRES	
EXPRATIO	EXPRATIO GQ	EXPRATIO GQ ST
IMP RAT EXP	IMP RAT EXP GQ	IMP RAT EXP GQ ST
MAXRATIO	MAXRATIO GQ	MAXRATIO GQ ST
IMP RAT MAX	IMP RAT MAX GQ	IMP RAT MAX GQ ST
CURRRATIO	CURRRATIO GQ	CURRATIO GQ ST
IMP RAT CURR	IMP RAT CURR GQ	IMP RAT CURR GQ ST
MAXCURRRATIO	MAXCURRRATIO GQ	MAXCURRRATIO GQ ST
IMP RAT MAXCURR	IMP RAT MAXCURR GQ	IMP RAT MAXCURR GQ ST
MEDGP	MEDGP GQ	MEDGP GQ ST
IMP GP	IMP FLAG	

Name this file gq_mafid_dssd_out.sas7bdat

Andrew Keller, Julianne Zamora, Tim Kennel
December ~~23~~24, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into six sections:

1. Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation
2. Running HB Edits
3. Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation
4. Creating Imputed Values
5. Apply Ordering to Select Final Imputed Value
6. Create Output File

Input Files:

1. [redacted].sas7bdat
2. [redacted].sas7bdat
[redacted].sas7bdat
~~3.~~4. CES 501 results
~~4. CES 301 results~~

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file [redacted].sas7bdat) referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.

Rename GQ_INITIAL_STATUS to GQ_PRE_STATUS.
Rename GQ_INITIAL_UNRES to GQ_PRE_UNRES.
Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file [redacted].sas7bdat), referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.

D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
   a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

Section 2: HB Edits

A. Calculate Ratios for editing.
   a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘) and GQ_INITIAL_POP >0, then
      i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT
      ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT
      iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE
      iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP
   b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.
   a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR
   b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100

3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100
7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

  a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

  b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

  c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

  d. For each MAFID, transform the ratio to create **SVALUE**.

    i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

    ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)

  e. For each MAFID, transform SVALUE to create **EVALUE**.

    i. EVALUE = SVALUE * max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}$^{0.5}$

    ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

  f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

    i. **E_Q1** = first quartile EVALUE

    ii. **E_MED** = median EVALUE

    iii. **E_Q3** = third quartile EVALUE

  g. For each GQTYPE, define upper and lower bounds.

    i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

    ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

    iii. **LOWER_C1** = E_MED – C1 * D_Q1

    iv. **LOWER_C2** = E_MED – C2 * D_Q1

    v. **LOWER_C3** = E_MED – C3 * D_Q1

    vi. **UPPER_C1** = E_MED + C1 * D_Q3

    vii. **UPPER_C2** = E_MED + C2 * D_Q3

    viii. **UPPER_C3** = E_MED + C3 * D_Q3

  h. For each MAFID, create **FLAG[X]**.

    i. If EVALUE is missing, FLAG[X] = 'M'

    ii. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = 'R'

    iii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'

    iv. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0 and FLAGA not in ('S','I') and FLAGB not in ('S','I') and FLAGC not in ('S','I') and FLAGD not in ('S','I'). Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

  i. If FLAG[X] = 'I' and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = 'S'.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

  a. If (FLAGA = 'I' or FLAGB = 'I' or FLAGC = 'I' or FLAGD = 'I') and IMPUTE_NEEDED ne 'N' then

    i. **GP = .**

    ii. **UNRES** = 1

  b. Otherwise,

    i. **GP =** GQ_INITIAL_POP

    ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

  a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values

    i. Calculate Ratios.

    We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):

      1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
      2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
      3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
      4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
      5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
      6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

    ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_EXP_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGA in (‘ ‘,’R’). Assign these values as **EXP_PERS_10** and **EXP_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_EXP_PERS_CNT.

1. Assign **EXP_PERS_TRUNC** = GQ_SIZE_EXP_PERS_CNT
2. If GQ_SIZE_EXP_PERS_CNT > EXP_PERS_90 then set **EXP_PERS_TRUNC** = EXP_PERS_90

~~6.~~3. If GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_EXP_PERS_CNT < EXP_PERS_10 then set **EXP_PERS_TRUNC** = EXP_PERS_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_EXP** = CEIL (~~GQ_SIZE_EXP_PERS_CNT~~EXP_PERS_TRUNC*EXPRATIO)
2. **IMP_RAT_EXP_GQ** = CEIL (~~GQ_SIZE_EXP_PERS_CNT~~EXP_PERS_TRUNC*EXPRATIO_GQ)
3. **IMP_RAT_EXP_GQ_ST** = CEIL (EXP_PERS_TRUNC~~GQ_SIZE_EXP_PERS_CNT~~*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagB in (‘’,’R’):

1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_MAX_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGB in (‘ ‘,’R’). Assign these values as **MAX_PERS_10** and **MAX_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_MAX_PERS_CNT.

1. Assign **MAX_PERS_TRUNC** = GQ_SIZE_MAX_PERS_CNT
2. If GQ_SIZE_MAX_PERS_CNT > MAX_PERS_90 then set **MAX_PERS_TRUNC** = MAX_PERS_90

~~7.~~3. If GQ_SIZE_MAX_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT < MAX_PERS_10 then set **MAX_PERS_TRUNC** = MAX_PERS_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_MAX** = CEIL (~~GQ_SIZE_MAX_PERS_CNT~~MAX_PERS_TRUNC*MAXRATIO)
2. **IMP_RAT_MAX_GQ** = CEIL (~~GQ_SIZE_MAX_PERS_CNT~~MAX_PERS_TRUNC*MAXRATIO_GQ)
3. **IMPRAT_MAX_GQ_ST** = CEIL (~~GQ_SIZE_MAX_PERS_CNT~~MAX_PERS_TRUNC*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):

1. Sum the GP and GQCURRSIZE value **for the nation.**
2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRSIZE for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGC in (' ','R'). Assign these values as **CURRSIZE_10** and **CURRSIZE_90** respectively. For each MAFID where UNRES = 1 , assign truncated values of GQCURRSIZE.

1. Assign **CURRSIZE_TRUNC** = GQCURRSIZE
2. If GQCURRSIZE > CURRSIZE_90 then set **CURRSIZE_TRUNC** = CURRSIZE_90
~~6.~~3. If GQCURRSIZE > 0 and GQCURRSIZE < CURRSIZE_10 then set **CURRSIZE_TRUNC** = CURRSIZE_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURR** = CEIL (CURRSIZE_TRUNC~~GQCURRSIZE~~*CURRSIZERATIO)
2. **IMP_RAT_CURR_GQ** = CEIL (CURRSIZE_TRUNC~~GQCURRSIZE~~*CURRSIZERATIO_GQ)
3. **IMP_RAT_CURR_GQ_ST** = CEIL (CURRSIZE_TRUNC~~GQCURRSIZE~~*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagD in (‘’,’R’):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR calculate the 10th and 90th percentiles of GQCURRMAXPOP~~SIZE~~ for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGD in (‘ ‘ ’R’). Assign these values as **CURRMAX_10** and **CURRMAX_90** respectively.

For each MAFID where UNRES = 1 assign truncated values of GQCURRMAXPOP.

1. Assign **CURRMAX_TRUNC** = GQCURRMAXPOP
2. If GQCURRMAXPOP > CURRMAX_90 then set **CURRMAX_TRUNC** = CURRMAX_90
3. If GQCURRMAXPOP > 0 and GQCURRMAXPOP < CURRMAX_10 then set **CURRMAX_TRUNC** = CURRMAX_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (~~GQCURRMAXPOP~~CURRMAX_TRUNC*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (CURRMAX_TRUNCT~~GQCURRMAXPOP~~*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (CURRMAX_TRUNC~~GQCURRMAXPOP~~*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’,’R’):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**

2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

iv. Assign values. For each MAFID, calculate the following values:

1. **IMP_MEDGP_GQ_ST** = CEIL(MEDGP_GQ_ST)
2. **IMP_MEDGP_GQ** = CEIL(MEDGP_GQ)
3. **IMP_MEDGP** = CEIL(MEDGP)

~~C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.~~

~~a. Define MAXPOP variable.~~

~~i. if GQCURRMAXPOP > 0 then **MAXPOP** = log(GQCURRMAXPOP);~~

~~ii. if GQCURRMAXPOP = 0 then **MAXPOP** = .;~~

~~b. Define the fitting universe (ratiofile) as this: FLAGA in (' ','R') and FLAGB in (' ','R') and FLAGC in (' ','R') and FLAGD in (' ','R') and unres = 0 and FOCS_ER_CB_CODE = ''~~

~~c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.~~

~~d. Fit and score this model:~~

```
~~proc genmod data = ratiofile;~~
~~     class gqtypcur;~~
~~     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /~~
~~           link = log d = poisson offset = maxpop maxiter = 500;~~
~~  store params;~~
~~     output out = poi_pred PREDICTED = pr_size;~~
~~run;~~

~~proc plm source=params;~~
~~  score data = nomaxscore out=nomaxscoreout/ ilink;~~
~~run;~~
```

~~e. Take the ceiling function of the predicted count. Call this **IMP_POISSON_COUNT.**~~

~~D.~~C. Fold in CES 501 results

Formatted: Font: Not Bold

Commented [JEZ(F1]: Remove?

Commented [JEZ(F2]: Residual Method

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table hierarchically as follows, if IMP_POISSON_COUNT is not missing, assign IMP_GP = IMP_POISSON_COUNT and assign IMP_FLAG = 201. If IMP_POISSON_COUNT is missing, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. Continue on through the table until all MAFIDs with a ~~in~~ unres = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	**IMP_FLAG**
~~IMP POISSON COUNT~~	~~201~~
IMP RAT EXP GQ ST	101
IMP RAT EXP GQ	102
IMP RAT EXP	103
IMP RAT MAX GQ ST	104
IMP RAT MAX GQ	105
IMP RAT MAX	106
IMP RAT CURR GQ ST	107
IMP RAT CURR GQ	108
‘IMP RAT CURR	109
IMP RAT CURRMAX GQ ST	110
IMP RAT CURRMAX GQ	111
IMP RAT CURRMAX	112
MEDGP GQ ST	401
MEDGP GQ	402
MEDGP	403

Section 6: Create Output File

Output GQ MAFID, adding the following variables:

MAFID		
FLAGA	FLAGB	
FLAGC	FLAGD	
GP	UNRES	
EXPRATIO	EXPRATIO GQ	EXPRATIO GQ ST
EXP PERS 10	EXP PERS 90	EXP PERS TRUNC
IMP RAT EXP	IMP RAT EXP GQ	IMP RAT EXP GQ ST
MAXRATIO	MAXRATIO GQ	MAXRATIO GQ ST
MAX PERS 10	MAX PERS 90	MAX PERS TRUNC
IMP RAT MAX	IMP RAT MAX GQ	IMP RAT MAX GQ ST
CURRRATIO	CURRRATIO GQ	CURRATIO GQ ST
CURRSIZE 10	CURRSIZE 90	CURRSIZE TRUNC
IMP RAT CURR	IMP RAT CURR GQ	IMP RAT CURR GQ ST
MAXCURRRATIO	MAXCURRRATIO GQ	MAXCURRRATIO GQ ST
CURRMAX 10	CURRMAX 90	CURRMAX TRUNC
IMP RAT ~~MAX~~CURRMAX	IMP RAT ~~MAX~~CURRMAX GQ	IMP RAT ~~MAX~~CURRMAX GQ ST
IMP MEDGP	IMP MEDGP GQ	IMP MEDGP GQ ST
IMP GP	IMP FLAG	
GQCURRMAXPOP		
GQCURRSIZE		
GQ SIZE EXP PERS CNT		
GQ SIZE MAX PERS CNT		

Name this file gq_mafid_dssd_out.sas7bdat

Andrew Keller, Julianne Zamora, Tim Kennel
December ~~23~~24, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into six sections:

1. Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation
2. Running HB Edits
3. Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation
4. Creating Imputed Values
5. Apply Ordering to Select Final Imputed Value
6. Create Output File

Input Files:

1. [redacted].sas7bdat
2. [redacted].sas7bdat
~~2.~~3 [redacted].sas7bdat
~~3.~~4. CES 501 results
~~4. CES 301 results~~

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file [redacted].sas7bdat) referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.

Rename GQ_INITIAL_STATUS to GQ_PRE_STATUS.
Rename GQ_INITIAL_UNRES to GQ_PRE_UNRES.
Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file [redacted].sas7bdat), referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.

D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
    a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

Section 2: HB Edits

A. Calculate Ratios for editing.
    a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in ('O','R',' ') and GQ_INITIAL_POP > 0, then
        i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT
        ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT
        iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE
        iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP
    b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.
    a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR
    b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100

3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100
7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

d. For each MAFID, transform the ratio to create **SVALUE**.

i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)

e. For each MAFID, transform SVALUE to create **EVALUE**.

i. EVALUE = SVALUE * max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}$^{0.5}$

ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

i. **E_Q1** = first quartile EVALUE

ii. **E_MED** = median EVALUE

iii. **E_Q3** = third quartile EVALUE

g. For each GQTYPE, define upper and lower bounds.

i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

iii. **LOWER_C1** = E_MED – C1 * D_Q1

iv. **LOWER_C2** = E_MED – C2 * D_Q1

v. **LOWER_C3** = E_MED – C3 * D_Q1

vi. **UPPER_C1** = E_MED + C1 * D_Q3

vii. **UPPER_C2** = E_MED + C2 * D_Q3

viii. **UPPER_C3** = E_MED + C3 * D_Q3

h. For each MAFID, create **FLAG[X]**.

i. If EVALUE is missing, FLAG[X] = 'M'

ii. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = 'R'

iii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'

iv. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0 and FLAGA not in ('S','I') and FLAGB not in ('S','I') and FLAGC not in ('S','I') and FLAGD not in ('S','I'). Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

   i. If FLAG[X] = 'I' and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = 'S'.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

a. If (FLAGA = 'I' or FLAGB = 'I' or FLAGC = 'I' or FLAGD = 'I') and IMPUTE_NEEDED ne 'N' then

   i. **GP = .**

   ii. **UNRES** = 1

b. Otherwise,

   i. **GP =** GQ_INITIAL_POP

   ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values

   i. Calculate Ratios.

   We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):

   1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
   2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
   3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
   4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
   5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
   6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

   ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_EXP_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGA in (‘ ‘,’R’). Assign these values as **EXP_PERS_10** and **EXP_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_EXP_PERS_CNT.

1. Assign **EXP_PERS_TRUNC** = GQ_SIZE_EXP_PERS_CNT
2. If GQ_SIZE_EXP_PERS_CNT > EXP_PERS_90 then set **EXP_PERS_TRUNC** = EXP_PERS_90
~~6.~~3. If GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_EXP_PERS_CNT < EXP_PERS_10 then set **EXP_PERS_TRUNC** = EXP_PERS_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_EXP** = CEIL (~~GQ_SIZE_EXP_PERS_CNT~~EXP_PERS_TRUNC*EXPRATIO)
2. **IMP_RAT_EXP_GQ** = CEIL (~~GQ_SIZE_EXP_PERS_CNT~~EXP_PERS_TRUNC*EXPRATIO_GQ)
3. **IMP_RAT_EXP_GQ_ST** = CEIL (EXP_PERS_TRUNC~~GQ_SIZE_EXP_PERS_CNT~~*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagB in (‘’,’R’):

1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_MAX_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGB in (‘ ‘,’R’). Assign these values as **MAX_PERS_10** and **MAX_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_MAX_PERS_CNT.

1. Assign **MAX_PERS_TRUNC** = GQ_SIZE_MAX_PERS_CNT
2. If GQ_SIZE_MAX_PERS_CNT > MAX_PERS_90 then set **MAX_PERS_TRUNC** = MAX_PERS_90

~~7.~~3. If GQ_SIZE_MAX_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT < MAX_PERS_10 then set **MAX_PERS_TRUNC** = MAX_PERS_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_MAX** = CEIL (~~GQ_SIZE_MAX_PERS_CNT~~MAX_PERS_TRUNC*MAXRATIO)
2. **IMP_RAT_MAX_GQ** = CEIL (~~GQ_SIZE_MAX_PERS_CNT~~MAX_PERS_TRUNC*MAXRATIO_GQ)
3. **IMPRAT_MAX_GQ_ST** = CEIL (~~GQ_SIZE_MAX_PERS_CNT~~MAX_PERS_TRUNC*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):

1. Sum the GP and GQCURRSIZE value **for the nation.**
2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRSIZE for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGC in (' ','R').

Assign these values as **CURRSIZE_10** and **CURRSIZE_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRSIZE.

1. Assign **CURRSIZE_TRUNC** = GQCURRSIZE
2. If GQCURRSIZE > CURRSIZE_90 then set **CURRSIZE_TRUNC** = CURRSIZE_90

~~6.~~3. If GQCURRSIZE > 0 and GQCURRSIZE < CURRSIZE_10 then set **CURRSIZE_TRUNC** = CURRSIZE_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURR** = CEIL (CURRSIZE_TRUNC~~GQCURRSIZE~~*CURRSIZERATIO)
2. **IMP_RAT_CURR_GQ** = CEIL (CURRSIZE_TRUNC~~GQCURRSIZE~~*CURRSIZERATIO_GQ)
3. **IMP_RAT_CURR_GQ_ST** = CEIL (CURRSIZE_TRUNC~~GQCURRSIZE~~*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR calculate the 10th and 90th percentiles of GQCURRMAXPOP~~SIZE~~ for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGD in (' ' 'R'). Assign these values as **CURRMAX_10** and **CURRMAX_90** respectively.

For each MAFID where UNRES = 1 assign truncated values of GQCURRMAXPOP.

1. Assign **CURRMAX_TRUNC** = GQCURRMAXPOP
2. If GQCURRMAXPOP > CURRMAX_90 then set **CURRMAX_TRUNC** = CURRMAX_90
3. If GQCURRMAXPOP > 0 and GQCURRMAXPOP < CURRMAX_10 then set **CURRMAX_TRUNC** = CURRMAX_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (~~GQCURRMAXPOP~~CURRMAX_TRUNC*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (CURRMAX_TRUNCT~~GQCURRMAXPOP~~*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (CURRMAX_TRUNC~~GQCURRMAXPOP~~*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**

2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

Formatted: Font: Not Bold

iv. Assign values. For each MAFID, calculate the following values:
1. **IMP_MEDGP_GQ_ST** = CEIL(MEDGP_GQ_ST)
2. **IMP_MEDGP_GQ** = CEIL(MEDGP_GQ)
3. **IMP_MEDGP** = CEIL(MEDGP)

~~C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.~~

~~a. Define MAXPOP variable.~~
~~i. if GQCURRMAXPOP > 0 then **MAXPOP** = log(GQCURRMAXPOP);~~
~~ii. if GQCURRMAXPOP = 0 then **MAXPOP** = .;~~

~~b. Define the fitting universe (ratiofile) as this: FLAGA in (' ','R') and FLAGB in (' ','R') and FLAGC in (' ','R') and FLAGD in (' ','R') and unres = 0 and FOCS_ER_CB_CODE = "~~

~~c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.~~

~~d. Fit and score this model:~~

```
~~proc genmod data = ratiofile;~~
~~     class gqtypcur;~~
~~     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT~~
~~GQ_SIZE_EXP_PERS_CNT /~~
~~           link = log d = poisson offset = maxpop maxiter = 500;~~
~~  store params;~~
~~     output out = poi_pred PREDICTED = pr_size;~~
~~run;~~

~~proc plm source=params;~~
~~  score data = nomaxscore out=nomaxscoreout/ ilink;~~
~~run;~~
```

~~e. Take the ceiling function of the predicted count. Call this **IMP_POISSON_COUNT.**~~

Commented [JEZ(F1)]: Remove?

C. Residual method: using a hybrid of the ratio imputes created in the previous step a percentile method based on Greek/non-Greek status and allocation of a facility-level residual to individual MAFIDs.

a. Ingest the file referred to as **MAFID_FRAT_SORO**
i. On this file **FLAG_GREEK_LETTER**=1 indicates that GQ has been identified as a fraternity or sorority house. Otherwise **FLAG_GREEK_LETTER**=0.

b. Ingest the file referred to as **UNITID_MAFID_LINKS.**
i. When reading in **UNITID_MAFID_LINKS**, keep only the variables **MAFID, UNITID, MATCH_STEP_NUM,** and **ROOMCAP.**
ii. Note: for records with **MATCH_STEP_NUM**=-1 **UNITID** will be missing.
iii. Note: for records with the same value of UNITID ROOMCAP will be the same.

c. Merge **MAFID_FRAT_SORO** and **UNITID_MAFID_LINKS** to ***GQ_MAFID***, merging on MAFID, and keeping only records that are in ***GQ_MAFID.***

i. Note: For records that match, this should be a 1-to-1 match (MAFID should be unique in each of the 3 datasets).

ii. Note: only records with GQCURTYP=501 in ***GQ_MAFID*** should match to either of the other 2 datasets.

d. Select the subset of the merged dataset from the previous step with GQCURTYP=501.

i. NOTE: In this spec we will refer to this subset of the data as **GQ_COUNTS_ROOMCAP_GREEK**. This is only an intermediate dataset which will be merged back to the **GQ_MAFID** dataset at the end of this section of the spec (section 5.D).

e. Using GQ_COUNTS_ROOM_CAP_GREEK and the ratio impute variables created in section 4.A, create a temporary impute variable IMP_GP_TEMP using the hierarchy shown in the following table. If IMP_RAT_EXP_GQ_ST is not missing assign IMP_GP_TEMP= IMP_RAT_EXP_GQ_ST and set ALREADY_IMPUTED=1. If IMP_RAT_EXP_GQ_ST is missing and IMP_RAT_EXP_GQ is not missing assign IMP_GP_TEMP= IMP_RAT_EXP_GQ and set ALREADY_IMPUTED=1. Continue through the table until all the variables in the table have been exhausted. For any remaining MAFIDs for which a value has not been assigned to IMP_GP_TEMP set ALREADY_IMPUTED=0;

IMP_GP_TEMP assignment hierarchy
IMP_RAT_EXP_GQ_ST
IMP_RAT_EXP_GQ
IMP_RAT_MAX_GQ_ST
IMP_RAT_MAX_GQ
IMP_RAT_CURR_GQ_ST
IMP_RAT_CURR_GQ
IMP_RAT_CURRMAX_GQ_ST
IMP_RAT_CURRMAX_GQ

f. Using only MAFIDs in **GQ_COUNTS_ROOMCAP_GREEK** with UNRES = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’,’R’), create 3 GP median variables and 3 GP maximum variables:

i. For each UNITID-FLAG_GREEK_LETTER combination with enough MAFIDs:

1. Calculate the median value of GP. Call this **P50_GP_UNIT_BY_GRK**
2. Calculate the maximum value of GP. Call this **MAX_GP_UNIT_BY_GRK.**
3. Merge the P50_GP_UNIT_BY_GRK and MAX_GP_UNIT_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on UNITID and FLAG_GREEK_LETTER.

ii. For each BCUSTATEFP-FLAG_GREEK_LETTER combination with enough MAFIDs:

1. Calculate the median value of GP. Call this **P50_GP_ST_BY_GRK**.
2. Calculate the maximum value of GP. Call this **MAX_GP_ST_BY_GRK**.
3. Merge P50_GP_ST_BY_GRK and MAX_GP_ST_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on BCUSTATEFP-FLAG_GREEK_LETTER combinations.

iii. For each value of FLAG_GREEK_LETTER:

      1. Calculate the median value of GP. Call this **P50_GP_BY_GRK.**
      2. Calculate the maximum value of GP. Call this **MAX_GP_BY_GRK**.
      3. Merge P50_GP_BY_GRK and MAX_BP_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK** merging on FLAG_GREEK_LETTER.

g. For MAFIDs for which UNRES=1 FLAG_GREEK_LETTER=1 and ALREADY_IMPUTED=0 assign median Greek imputes to IMP_GP_TEMP and create up to 3 new impute variables using the following hierarchy:
   i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:
      1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
      2. Set ALREADY_IMPUTED=1
      3. Assign **MEDGP_GRK_UNIT**= IMP_GP_TEMP
   ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:
      1. assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
      2. set ALREADY_IMPUTED=1
      3. Assign **MEDGP_GRK_ST**= IMP_GP_TEMP
   iii. Otherwise:
      1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
      2. Set ALREADY_IMPUTED=1
      3. Assign **MEDGP_GRK**=IMP_GP_TEMP

h. Using **GQ_COUNTS_ROOMCAP_GREEK** by UNITID create unit-level sum variables (where a unit corresponds to a single UNITID, which corresponds to a single a university or college)
   i. Create unit-level sums (i.e. by UNITID) of GQCURRMAXPOP using only observations where flagD in (‘’ ’R’). Note: these are the “good” values of GQCURRMAXPOP. Note that for this sum we don’t care what the value of GP is even it is a true 0. We are just trying to come up with a maximum number of people that these GQs *could* house, so that we can subtract the sum from the college-level IPEDS ROOMCAP variable. For reference later in the spec call this sum **UNIT_MAXPOP_SUM**.

   ii. Using only the GQs with unres=0 and flagD **not** in (‘’,’R’), by UNITID, create unit-level sums of GP. Call this sum **UNIT_2020POP_SUM**.
   iii. Using only the GQs with unres=1 and flagD **not** in (‘’,’R’), by UNITID, create unit-level sums of IMP_GP_TEMP. Call this **UNIT_POP_IMPUTED_SUM**.
   iv. Create **UNIT_CAP_SUM** = the unit-level sum of UNIT_MAXPOP_SUM UNIT_2020POP_SUM and UNIT_POP_IMPUTED_SUM

i. For each MAFID, calculate UNIT_RESIDUAL = ROOMCAP – UNIT_CAP_SUM (this will be the same value for MAFIDs with the same UNITID)

j. For each MAFID with UNIT_RESIDUAL<=0 UNRES=1 and ALREADY_IMPUTED=0 assign values to IMP_GP_TEMP and create 3 new (non-Greek) median impute variables using the following hierarchy:
   i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:
      1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
      2. Set ALREADY_IMPUTED=1
      3. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP
   ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:

```
            1. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
            2. Set ALREADY_IMPUTED=1
            3. Assign MEDGP_nonGRK_ST= IMP_GP_TEMP
        iii. Otherwise:
            1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
            2. Set ALREADY_IMPUTED=1
            3. Assign MEDGP_nonGRK=IMP_GP_TEMP
k. For each (non-missing) UNITID with UNIT_RESIDUAL>0, count the MAFIDs associated with that UNITID that have UNRES=1 and ALREADY_IMPUTED=0. Call this count UNIT_RESID_GQ_COUNT.
l. For MAFIDs with UNIT_RESIDUAL>0 UNIT_RESID_GQ_COUNT=1 UNRES=1 and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy:
    i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK then assign values to IMP_GP_TEMP using the following sub-hierarchy:
        1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:
            a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
            b. Set ALREADY_IMPUTED=1
            c. Assign MEDGP_nonGRK_UNIT= IMP_GP_TEMP
        2. Otherwise (i.e. if P50_GP_UNIT_BY_GRK<=0 or missing) if MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing, then:
            a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
            b. Set ALREADY_IMPUTED=1
            c. Assign MEDGP_nonGRK_ST= IMP_GP_TEMP
        3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:
            a. Assign IMP_GP_TEMP= P50_GP_BY_GRK
            b. Set ALREADY_IMPUTED=1
            c. Assign MEDGP_nonGRK=IMP_GP_TEMP
    ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL < MAX_GP_UNIT_BY_GRK, then assign values as follows:
        1. Assign IMP_GP_TEMP=UNIT_RESIDUAL
        2. Set ALREADY_IMPUTED=1
        3. Assign IMP_RESID_1GQ=IMP_GP_TEMP
m. For MAFIDs with UNIT_RESIDUAL>0 UNIT_RESID_GQ_COUNT>1 UNRES=1 and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy. (NOTE: steps i.1-i.3 are the same as steps i.1-i.3 in step l above):
    i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK then assign values to IMP_GP_TEMP using the following sub-hierarchy:
        1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:
            a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
            b. Set ALREADY_IMPUTED=1
            c. Assign MEDGP_nonGRK_UNIT= IMP_GP_TEMP
```

            2. Otherwise (i.e., if P50 GP UNIT BY GRK<=0 or missing), if MAX GP ST BY GRK>0 and non-missing and UNIT RESIDUAL > MAX GP ST BY GRK and P50 GP ST BY GRK>0 and non-missing, then:
                a. Assign IMP GP TEMP= P50 GP ST BY GRK
                b. Set ALREADY IMPUTED=1
                c. Assign **MEDGP nonGRK ST**= IMP GP TEMP
            3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:
                a. Assign IMP GP TEMP= P50 GP BY GRK
                b. Set ALREADY IMPUTED=1
                c. Assign **MEDGP nonGRK**=IMP GP TEMP
        ii. If MAX GP UNIT BY GRK=0 or missing or UNIT RESIDUAL < MAX GP UNIT BY GRK, then assign values as follows:
            1. Assign IMP GP TEMP=UNIT RESIDUAL/UNIT RESID GQ COUNT
            2. Set ALREADY IMPUTED=1
            3. Assign **IMP RESID NGQ**=IMP GP TEMP
    n. Do a cross-tabulation of the variables UNRES and ALREADY IMPUTED. If ALREADY IMPUTED is always 1 when UNRES=1, then imputations have been calculated for all MAFIDS with GQCURTYP 501.
    o. Keep the variables **MEDGP GRK UNIT, MEDGP GRK ST, MEDGP GRK, MEDGP nonGRK UNIT, MEDGP nonGRK ST, MEDGP nonGRK, IMP RESID 1GQ** and **IMP RESID NGQ.** Drop all other variables created in this section

~~D. Fold in CES 501 results~~

Commented [JEZ(F2]: Residual Method

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table hierarchically as follows, if IMP_POISSON_COUNT is not missing, assign IMP_GP = IMP_POISSON_COUNT and assign IMP_FLAG = 201. If IMP_POISSON_COUNT is missing, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. Continue on through the table until all MAFIDs with a ~~in~~ unres = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	**IMP_FLAG**
~~IMP POISSON COUNT~~	~~201~~
IMP RAT EXP GQ ST	101
IMP RAT EXP GQ	102
IMP RAT EXP	103
IMP RAT MAX GQ ST	104
IMP RAT MAX GQ	105
IMP RAT MAX	106
IMP RAT CURR GQ ST	107
IMP RAT CURR GQ	108
‘IMP RAT CURR	109

IMP RAT CURRMAX GQ ST	110
IMP RAT CURRMAX GQ	111
IMP RAT CURRMAX	112
MEDGP GRK UNIT	301
MEDGP GRK ST	302
MEDGP GRK	303
MEDGP nonGRK UNIT	304
MEDGP nonGRK ST	305
MEDGP nonGRK	306
IMP RESID 1GQ	307
IMP RESID NGQ	308
MEDGP GQ ST	401
MEDGP GQ	402
MEDGP	403

Section 6: Create Output File

Output GQ MAFID, adding the following variables:

MAFID		
FLAGA	FLAGB	
FLAGC	FLAGD	
GP	UNRES	
EXPRATIO	EXPRATIO GQ	EXPRATIO GQ ST
EXP PERS 10	EXP PERS 90	EXP PERS TRUNC
IMP RAT EXP	IMP RAT EXP GQ	IMP RAT EXP GQ ST
MAXRATIO	MAXRATIO GQ	MAXRATIO GQ ST
MAX PERS 10	MAX PERS 90	MAX PERS TRUNC
IMP RAT MAX	IMP RAT MAX GQ	IMP RAT MAX GQ ST
CURRRATIO	CURRRATIO GQ	CURRATIO GQ ST
CURRSIZE 10	CURRSIZE 90	CURRSIZE TRUNC
IMP RAT CURR	IMP RAT CURR GQ	IMP RAT CURR GQ ST
MAXCURRRATIO	MAXCURRRATIO GQ	MAXCURRRATIO GQ ST
CURRMAX 10	CURRMAX 90	CURRMAX TRUNC
IMP RAT ~~MAX~~CURRMAX	IMP RAT ~~MAX~~CURRMAX GQ	IMP RAT ~~MAX~~CURRMAX GQ ST
IMP MEDGP	IMP MEDGP GQ	IMP MEDGP GQ ST
IMP GP	IMP FLAG	
GQCURRMAXPOP		
GQCURRSIZE		
GQ SIZE EXP PERS CNT		
GQ SIZE MAX PERS CNT		

Name this file gq_mafid_dssd_out.sas7bdat

Andrew Keller, Julianne Zamora, Tim Kennel, Kirk White
December 26, 2020
2020 Census Specification For Group Quarters Imputation

Introduction
The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into six sections:

1. Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation
2. Running HB Edits
3. Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation
4. Creating Imputed Values
5. Apply Ordering to Select Final Imputed Value
6. Create Output File

Input Files:

1. [redacted].sas7bdat (GQ_MAFID)
2. [redacted].sas7bdat (HBPARM)
3. [redacted].sas7bdat (GQ_DUP_MAFID)
4. [redacted].csv (MAFID_FRAT_SORO)
5. [redacted].sas7bdat (UNITID_MAFID_LINKS)

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation
This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file [redacted].sas7bdat), referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.
D. Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file [redacted].sas7bdat), referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.
D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
    a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

Section 2: HB Edits

A. Calculate Ratios for editing.

a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘) and GQ_INITIAL_POP > 0, then

i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT

ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT

iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE

iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP

b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.

a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR

b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* (hbparm.sas7bdat) file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175
4	D	25	50	100

5	D	100	125	175
6	D	25	50	100
7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

   a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

   b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

   c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

   d. For each MAFID, transform the ratio to create **SVALUE**.

      i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

      ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO) - 1

   e. For each MAFID, transform SVALUE to create **EVALUE**.

      i. Calculate MAX_INTIAL_POP as max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}

      ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

      iii. EVALUE = SVALUE *(MAX_INITIAL_POP) $^{1/2}$

   f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

      i. **E_Q1** = first quartile EVALUE

      ii. **E_MED** = median EVALUE

      iii. **E_Q3** = third quartile EVALUE

   g. For each GQTYPE, define upper and lower bounds.

      i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

      ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

      iii. **LOWER_C1** = E_MED – C1 * D_Q1

      iv. **LOWER_C2** = E_MED – C2 * D_Q1

      v. **LOWER_C3** = E_MED – C3 * D_Q1

      vi. **UPPER_C1** = E_MED + C1 * D_Q3

      vii. **UPPER_C2** = E_MED + C2 * D_Q3

      viii. **UPPER_C3** = E_MED + C3 * D_Q3

   h. For each MAFID, create **FLAG[X]**.

      i. If EVALUE is missing, FLAG[X] = 'M'

      ii. Otherwise, apply the following conditions, without nesting (i.e. apply each 'if' statement separately).

         1. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = 'R'

2. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'
3. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0 and FLAGA not in ('S','I') and FLAGB not in ('S','I') and FLAGC not in ('S','I') and FLAGD not in ('S','I') and GQ_INITIAL_POP > 0. Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

i. If FLAG[X] = 'I' and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = 'S'.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

a. If (FLAGA = 'I' or FLAGB = 'I' or FLAGC = 'I' or FLAGD = 'I') and IMPUTE_NEEDED ne 'N' then

i. **GP = .**

ii. **UNRES** = 1

b. Otherwise,

i. **GP =** GQ_INITIAL_POP

ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):

1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**

        4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
        5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
        6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

    ii. Calculate Bounds.

    For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_EXP_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGA in (‘ ‘,’R’). Assign these values as **EXP_PERS_10** and **EXP_PERS_90** respectively.

    For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_EXP_PERS_CNT.

        1. Assign **EXP_PERS_TRUNC** = GQ_SIZE_EXP_PERS_CNT
        2. If GQ_SIZE_EXP_PERS_CNT > EXP_PERS_90 then set **EXP_PERS_TRUNC** = EXP_PERS_90
        3. If GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_EXP_PERS_CNT < EXP_PERS_10 then set **EXP_PERS_TRUNC** = EXP_PERS_10.

    iii. Assign values. For each MAFID, calculate the following values:

        1. **IMP_RAT_EXP** = CEIL (EXP_PERS_TRUNC*EXPRATIO)
        2. **IMP_RAT_EXP_GQ** = CEIL (EXP_PERS_TRUNC*EXPRATIO_GQ)
        3. **IMP_RAT_EXP_GQ_ST** = CEIL (EXP_PERS_TRUNC*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values

    i. Calculate Ratios.

    We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagB in (‘’,’R’):

        1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
        2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
        3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
        4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
        5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
        6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

    ii. Calculate Bounds.

    For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_MAX_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGB in (‘ ‘,’R’). Assign these values as **MAX_PERS_10** and **MAX_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_MAX_PERS_CNT.

1. Assign **MAX_PERS_TRUNC** = GQ_SIZE_MAX_PERS_CNT
2. If GQ_SIZE_MAX_PERS_CNT > MAX_PERS_90 then set **MAX_PERS_TRUNC** = MAX_PERS_90
3. If GQ_SIZE_MAX_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT < MAX_PERS_10 then set **MAX_PERS_TRUNC** = MAX_PERS_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_MAX** = CEIL (MAX_PERS_TRUNC*MAXRATIO)
2. **IMP_RAT_MAX_GQ** = CEIL (MAX_PERS_TRUNC*MAXRATIO_GQ)
3. **IMPRAT_MAX_GQ_ST** = CEIL (MAX_PERS_TRUNC*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):

1. Sum the GP and GQCURRSIZE value **for the nation.**
2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRSIZE for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGC in (' ','R'). Assign these values as **CURRSIZE_10** and **CURRSIZE_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRSIZE.

1. Assign **CURRSIZE_TRUNC** = GQCURRSIZE
2. If GQCURRSIZE > CURRSIZE_90 then set **CURRSIZE_TRUNC** = CURRSIZE_90
3. If GQCURRSIZE > 0 and GQCURRSIZE < CURRSIZE_10 then set **CURRSIZE_TRUNC** = CURRSIZE_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURR** = CEIL (CURRSIZE_TRUNC*CURRSIZERATIO)
2. **IMP_RAT_CURR_GQ** = CEIL (CURRSIZE_TRUNC*CURRSIZERATIO_GQ)
3. **IMP_RAT_CURR_GQ_ST** = CEIL (CURRSIZE_TRUNC*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRMAXPOP for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGD in (' ','R'). Assign these values as **CURRMAX_10** and **CURRMAX_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRMAXPOP.

1. Assign **CURRMAX_TRUNC** = GQCURRMAXPOP
2. If GQCURRMAXPOP > CURRMAX_90 then set **CURRMAX_TRUNC** = CURRMAX_90
3. If GQCURRMAXPOP > 0 and GQCURRMAXPOP < CURRMAX_10 then set **CURRMAX_TRUNC** = CURRMAX_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (CURRMAX_TRUNC*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (CURRMAX_TRUNC*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (CURRMAX_TRUNC*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**

2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

iv. Assign values. For each MAFID, calculate the following values:

1. **IMP_MEDGP_GQ_ST** = CEIL(MEDGP_GQ_ST)
2. **IMP_MEDGP_GQ** = CEIL(MEDGP_GQ)
3. **IMP_MEDGP** = CEIL(MEDGP)

C. CES method: impute using a hybrid of the ratio imputes created in the previous step, a percentile method based on Greek/non-Greek status, and a facility-level residual allocation method.

a. Ingest the file referred to as **MAFID_FRAT_SORO**

i. On this file **FLAG_GREEK_LETTER**=1 indicates that GQ has been identified as a fraternity or sorority house. Otherwise **FLAG_GREEK_LETTER**=0.

b. Ingest the file referred to as **UNITID_MAFID_LINKS**.

i. When reading in **UNITID_MAFID_LINKS,** keep only the variables **MAFID, UNITID, MATCH_STEP_NUM** and **ROOMCAP.**

ii. Note: for records with **MATCH_STEP_NUM**=-1 **UNITID** will be missing.

iii. Note: for records with the same value of UNITID, ROOMCAP will be the same.

c. Merge **MAFID_FRAT_SORO** and **UNITID_MAFID_LINKS** to ***GQ_MAFID***, merging on MAFID and keeping only records that are in ***GQ_MAFID.***

i. Note: For records that match this should be a 1-to-1 match (MAFID should be unique in each of the 3 datasets).

ii. Note: only records with GQCURTYP=501 in ***GQ_MAFID*** should match to either of the other 2 datasets.

d. Select the subset of the merged dataset from the previous step with GQCURTYP=501.

i. NOTE: In this spec we will refer to this subset of the data as **GQ_COUNTS_ROOMCAP_GREEK**. This is only an intermediate dataset which will be merged back to the **GQ_MAFID** dataset at the end of this section of the spec (section 5.D).

e. Using GQ_COUNTS_ROOM_CAP_GREEK and the ratio impute variables created in section 4.A create a temporary impute variable IMP_GP_TEMP using the hierarchy shown in the following table. If IMP_RAT_EXP_GQ_ST is not missing assign IMP_GP_TEMP= IMP_RAT_EXP_GQ_ST and set ALREADY_IMPUTED=1. If IMP_RAT_EXP_GQ_ST is missing and IMP_RAT_EXP_GQ is not missing, assign IMP_GP_TEMP= IMP_RAT_EXP_GQ and set ALREADY_IMPUTED=1. Continue through the table until all the variables in the table have been exhausted. For any remaining MAFIDs for which a value has not been assigned to IMP_GP_TEMP set ALREADY_IMPUTED=0;

IMP_GP_TEMP assignment hierarchy
IMP_RAT_EXP_GQ_ST
IMP_RAT_EXP_GQ
IMP_RAT_MAX_GQ_ST
IMP_RAT_MAX_GQ

IMP_RAT_CURR_GQ_ST
IMP_RAT_CURR_GQ
IMP_RAT_CURRMAX_GQ_ST
IMP_RAT_CURRMAX_GQ

f. Using only MAFIDs in **GQ_COUNTS_ROOMCAP_GREEK** with UNRES = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’ ’R’) create 3 GP median variables and 3 GP maximum variables:

  i. For each UNITID-FLAG_GREEK_LETTER combination MAFIDs:

    1. Calculate the median value of GP. Call this **P50_GP_UNIT_BY_GRK**
    2. Calculate the maximum value of GP. Call this **MAX_GP_UNIT_BY_GRK.**
    3. Merge the P50_GP_UNIT_BY_GRK and MAX_GP_UNIT_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK** merging on UNITID and FLAG_GREEK_LETTER.
    4. Note these values will be missing if there are not enough observations for the UNITID-FLAG_GREEK_LETTER combination.

  ~~v.~~ii. For each BCUSTATEFP-FLAG_GREEK_LETTER combination ~~with enough~~ MAFIDs:

    1. Calculate the median value of GP. Call this **P50_GP_ST_BY_GRK**.
    2. Calculate the maximum value of GP. Call this **MAX_GP_ST_BY_GRK**.
    3. Merge P50_GP_ST_BY_GRK and MAX_GP_ST_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on BCUSTATEFP-FLAG_GREEK_LETTER combinations.
    4. Note, these values will be missing if there are not enough observations for the BCUSTATEFP-FLAG_GREEK_LETTER combination.

  iii. For each value of FLAG_GREEK_LETTER:

    1. Calculate the median value of GP. Call this **P50_GP_BY_GRK.**
    2. Calculate the maximum value of GP. Call this **MAX_GP_BY_GRK**.
    3. Merge P50_GP_BY_GRK and MAX_BP_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK** merging on FLAG_GREEK_LETTER.

g. For MAFIDs for which UNRES=1 FLAG_GREEK_LETTER=1 and ALREADY_IMPUTED=0 assign median Greek imputes to IMP_GP_TEMP and create up to 3 new impute variables using the following hierarchy:

  i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:

    1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
    2. Set ALREADY_IMPUTED=1
    3. Assign **MEDGP_GRK_UNIT**= IMP_GP_TEMP

  ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:

    1. assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
    2. set ALREADY_IMPUTED=1
    3. Assign **MEDGP_GRK_ST**= IMP_GP_TEMP

  iii. Otherwise:

    1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
    2. Set ALREADY_IMPUTED=1
    3. Assign **MEDGP_GRK**=IMP_GP_TEMP

h. Using **GQ_COUNTS_ROOMCAP_GREEK**, by UNITID, create unit-level sum variables (where a unit corresponds to a single UNITID, which corresponds to a single a university or college)

   i. Create unit-level sums (i.e. by UNITID) of GQCURRMAXPOP using only observations where flagD in (" 'R'). Note: these are the "good" values of GQCURRMAXPOP. Note that for this sum, we don't care what the value of GP is, even it is a true 0. We are just trying to come up with a maximum number of people that these GQs *could* house, so that we can subtract the sum from the college-level IPEDS ROOMCAP variable. For reference later in the spec call this sum **UNIT_MAXPOP_SUM.**

   ii. Using only the GQs with unres=0~~ and flagD **not** in (",'R')~~ and flagA not in ('I','S') and flagB not in ('I','S') and flagC not in ('I','S') and flagD = 'M' and GQCURRMAXPOP=. by UNITID create unit-level sums of GP. Call this sum **UNIT_2020POP_SUM.**

   iii. Using only the GQs with~~ unres=1 and flagD **not** in (" 'R')~~ (unres=1 or flagA = 'I' or flagB='I' or flagC='I' or flagD='I') and already_imputed=1 and GQCURRMAXPOP=., by UNITID, create unit-level sums of IMP_GP_TEMP. Call this **UNIT_POP_IMPUTED_SUM.**

   iv. Create **UNIT_CAP_SUM** = the unit-level sum of UNIT_MAXPOP_SUM UNIT_2020POP_SUM and UNIT_POP_IMPUTED_SUM

i. For each MAFID, calculate UNIT_RESIDUAL = ROOMCAP – UNIT_CAP_SUM (this will be the same value for MAFIDs with the same UNITID)

j. For each MAFID with UNIT_RESIDUAL<=0 UNRES=1 and ALREADY_IMPUTED=0 assign values to IMP_GP_TEMP and create 3 new (non-Greek) median impute variables using the following hierarchy:

   i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:
      1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
      2. Set ALREADY_IMPUTED=1
      3. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP

   ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:
      1. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
      2. Set ALREADY_IMPUTED=1
      3. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP

   iii. Otherwise:
      1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
      2. Set ALREADY_IMPUTED=1
      3. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

k. For each (non-missing) UNITID with UNIT_RESIDUAL>0 count the MAFIDs associated with that UNITID that have UNRES=1 and ALREADY_IMPUTED=0. Call this count UNIT_RESID_GQ_COUNT.

l. For MAFIDs with UNIT_RESIDUAL>0, UNIT_RESID_GQ_COUNT=1, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy:

   i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK then assign values to IMP_GP_TEMP using the following sub-hierarchy:

Commented [JEZ(F1]: Ask Kirk. This is like, if GQCURRMAXPOP is good, take that. Then if it's flagged but the GQ is resolved, take GP (so this includes suppressed). Then if it's unresolved, take imputed value. Some all of these to get a POP for the unit?

Commented [JEZ(F2R1]: This might change.

```
            1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:
                a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
                b. Set ALREADY_IMPUTED=1
                c. Assign MEDGP_nonGRK_UNIT= IMP_GP_TEMP
            2. Otherwise (i.e. if P50_GP_UNIT_BY_GRK<=0 or missing) if
               MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL >
               MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing,
               then:
                a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
                b. Set ALREADY_IMPUTED=1
                c. Assign MEDGP_nonGRK_ST= IMP_GP_TEMP
            3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:
                a. Assign IMP_GP_TEMP= P50_GP_BY_GRK
                b. Set ALREADY_IMPUTED=1
                c. Assign MEDGP_nonGRK=IMP_GP_TEMP
        ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL <
            MAX_GP_UNIT_BY_GRK, then assign values as follows:
            1. Assign IMP_GP_TEMP=UNIT_RESIDUAL
            2. Set ALREADY_IMPUTED=1
            3. Assign IMP_RESID_1GQ=IMP_GP_TEMP
m. For MAFIDs with UNIT_RESIDUAL>0 UNIT_RESID_GQ_COUNT>1 UNRES=1 and
   ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and
   create (up to) 1 new impute variables using the following hierarchy. (NOTE: steps i.1-i.3
   are the same as steps i.1-i.3 in step l above):
        i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL >
           MAX_GP_UNIT_BY_GRK then assign values to IMP_GP_TEMP using the
           following sub-hierarchy:
            1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:
                a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
                b. Set ALREADY_IMPUTED=1
                c. Assign MEDGP_nonGRK_UNIT= IMP_GP_TEMP
            2. Otherwise (i.e., if P50_GP_UNIT_BY_GRK<=0 or missing), if
               MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL >
               MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing,
               then:
                a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
                b. Set ALREADY_IMPUTED=1
                c. Assign MEDGP_nonGRK_ST= IMP_GP_TEMP
            3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:
                a. Assign IMP_GP_TEMP= P50_GP_BY_GRK
                b. Set ALREADY_IMPUTED=1
                c. Assign MEDGP_nonGRK=IMP_GP_TEMP
        ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL <
            MAX_GP_UNIT_BY_GRK, then assign values as follows:
            1. Assign IMP_GP_TEMP=UNIT_RESIDUAL/UNIT_RESID_GQ_COUNT
            2. Set ALREADY_IMPUTED=1
            3. Assign IMP_RESID_NGQ=IMP_GP_TEMP
```

n. Do a cross-tabulation of the variables UNRES and ALREADY IMPUTED. If ALREADY IMPUTED is always 1 when UNRES=1, then imputations have been calculated for all MAFIDS with GQCURTYP 501.

o. Keep the variables **MEDGP GRK UNIT, MEDGP GRK ST, MEDGP GRK, MEDGP nonGRK UNIT, MEDGP nonGRK ST, MEDGP nonGRK, IMP RESID 1GQ** and **IMP RESID NGQ.** Drop all other variables created in this section

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table hierarchically as follows, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. If IMP_RAT_EXP_GQ_ST is missing, if is not missing, assign IMP_GP = IMP_RAT_EXP_GQ and assign IMP_FLAG = 102. Continue on through the table until all MAFIDs with UNRES = 1 have a value for IMP_GP and IMP_FLAG.

IMP GP	IMP FLAG
IMP RAT EXP GQ ST	101
IMP RAT EXP GQ	102
IMP RAT EXP	103
IMP RAT MAX GQ ST	104
IMP RAT MAX GQ	105
IMP RAT MAX	106
IMP RAT CURR GQ ST	107
IMP RAT CURR GQ	108
IMP RAT CURR	109
IMP RAT CURRMAX GQ ST	110
IMP RAT CURRMAX GQ	111
IMP RAT CURRMAX	112
MEDGP GRK UNIT	301
MEDGP GRK ST	302
MEDGP GRK	303
MEDGP nonGRK UNIT	304
MEDGP nonGRK ST	305
MEDGP nonGRK	306
IMP RESID 1GQ	307
IMP RESID NGQ	308
IMP MEDGP GQ ST	401
IMP MEDGP GQ	402
IMP MEDGP	403

Section 6: Create Output Files

Output the following variables from GQMAFID:

Commented [JEZ(F3]: Ryan's recent files don't have geography on them...

MAFID	ACOCE	BCUCOUNTYFP
BCUSTATEFP	FACTLNAME	GQ SIZE EXP PERS CNT

GQ_SIZE_MAX_PERS_CNT	GQCONTACT	GQCURRMAXPOP
GQCURRSIZE	GQNAME	GQTYPCUR
GQ_INITIAL_STATUS	GQ_INITIAL_UNRES	GQ_INITIAL_POP
IMPUTE_NEEDED	FLAGA	FLAGB
FLAGC	FLAGD	GP
UNRES	IMP_GP	IMP_FLAG
EXPRATIO	EXPRATIO_GQ	EXPRATIO_GQ_ST
EXP_PERS_10	EXP_PERS_90	EXP_PERS_TRUNC
IMP_RAT_EXP	IMP_RAT_EXP_GQ	IMP_RAT_EXP_GQ_ST
MAXRATIO	MAXRATIO_GQ	MAXRATIO_GQ_ST
MAX_PERS_10	MAX_PERS_90	MAX_PERS_TRUNC
IMP_RAT_MAX	IMP_RAT_MAX_GQ	IMP_RAT_MAX_GQ_ST
CURRRATIO	CURRRATIO_GQ	CURRATIO_GQ_ST
CURRSIZE_10	CURRSIZE_90	CURRSIZE_TRUNC
IMP_RAT_CURR	IMP_RAT_CURR_GQ	IMP_RAT_CURR_GQ_ST
MAXCURRRATIO	MAXCURRRATIO_GQ	MAXCURRRATIO_GQ_ST
CURRMAX_10	CURRMAX_90	CURRMAX_TRUNC
IMP_RAT_CURRMAX	IMP_RAT_CURRMAX_GQ	IMP_RAT_CURRMAX_GQ_ST
MEDGP	MEDGP_GQ	MEDGP_GQ_ST
IMP_MEDGP	IMP_MEDGP_GQ	IMP_MEDGP_GQ_ST
MEDGP_GRK_UNIT	MEDGP_GRK_ST	MEDGP_GRK
MED_GP_nonGRK_UNIT	MEDGP_nonGRK_ST	MEDGP_nonGRK
IMP_RESID1GQ	IMP_RESID_NGQ	

Name this file gq_mafid_dssd_out_validation.sas7bdat

Output the following variables from GQMAFID:

MAFID	ACOCE	BCUCOUNTYFP
BCUSTATEFP	FACTLNAME	GQ_SIZE_EXP_PERS_CNT
GQ_SIZE_MAX_PERS_CNT	GQCONTACT	GQCURRMAXPOP
GQCURRSIZE	GQNAME	GQTYPCUR
GQ_INITIAL_STATUS	GQ_INITIAL_UNRES	GQ_INITIAL_POP
IMPUTE_NEEDED	FLAGA	FLAGB
FLAGC	FLAGD	GP
UNRES	IMP_GP	IMP_FLAG

Name this file gq_mafid_dssd_out_pop.sas7bdat. See POP data dictionary.

Andrew Keller, Julianne Zamora, Tim Kennel, Kirk White
December 26, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into six sections:

1. Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation
2. Running HB Edits
3. Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation
4. Creating Imputed Values
5. Apply Ordering to Select Final Imputed Value
6. Create Output File

Input Files:

1. [redacted].sas7bdat (GQ_MAFID)
2. [redacted].sas7bdat (HBPARM)
3. [redacted].sas7bdat (GQ_DUP_MAFID)
4. [redacted].csv (MAFID_FRAT_SORO)
5. [redacted].sas7bdat (UNITID_MAFID_LINKS)

Output Files: DSSD GQ Imputation Validation File (gq_mafid dssd out validation.sas7bdat)
DSSD GQ Imputation Review File for POP (gq_mafid_dssd_out_pop.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file [redacted].sas7bdat), referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.
D. Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file [redacted].sas7bdat), referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.
D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
    a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

Section 2: HB Edits

A. Calculate Ratios for editing.
   a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘) and GQ_INITIAL_POP > 0, then
      i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT
      ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT
      iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE
      iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP
   b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.
   a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR
   b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* (hbparm.sas7bdat) file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175

4	D	25	50	100
5	D	100	125	175
6	D	25	50	100
7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

d. For each MAFID, transform the ratio to create **SVALUE**.

i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO) - 1

e. For each MAFID, transform SVALUE to create **EVALUE**.

i. Calculate MAX_INTIAL_POP as max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}

ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

iii. EVALUE = SVALUE *(MAX_INITIAL_POP) $^{1/2}$

f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

i. **E_Q1** = first quartile EVALUE

ii. **E_MED** = median EVALUE

iii. **E_Q3** = third quartile EVALUE

g. For each GQTYPE, define upper and lower bounds.

i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

iii. **LOWER_C1** = E_MED – C1 * D_Q1

iv. **LOWER_C2** = E_MED – C2 * D_Q1

v. **LOWER_C3** = E_MED – C3 * D_Q1

vi. **UPPER_C1** = E_MED + C1 * D_Q3

vii. **UPPER_C2** = E_MED + C2 * D_Q3

viii. **UPPER_C3** = E_MED + C3 * D_Q3

h. For each MAFID, create **FLAG[X]**.

i. If EVALUE is missing, FLAG[X] = 'M'

ii. Otherwise, apply the following conditions, without nesting (i.e. apply each 'if' statement separately).

1. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = 'R'

2. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'
3. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0 and FLAGA not in ('S','I') and FLAGB not in ('S','I') and FLAGC not in ('S','I') and FLAGD not in ('S','I') and GQ_INITIAL_POP > 0. Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

i. If FLAG[X] = 'I' and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = 'S'.

c. For each MAFID, make the following updates:

i. If FLAGA = ' ' and FLAGB = 'I' then:

1. Set FLAGB = 'S'
2. If FLAGC = 'I' then set FLAGC = 'S'.
3. If FLAGD = 'I' then set FLAGD = 'S'.

ii. If FLAGA = ' ' and FLAGB = ' ' and FLAGC = 'I' then set FLAGC = 'S'.

~~i.~~iii. If FLAGA = ' ' and FLAGB = ' ' and FLAGC = ' ' and FLAGD = 'I' then set FLAGD = 'S'.

Commented [JEZ(F1]: Issue #2 with FLAGA and FLAGB = ' ' and FLAGC = 'I'.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

a. If (FLAGA = 'I' or FLAGB = 'I' or FLAGC = 'I' or FLAGD = 'I') and IMPUTE_NEEDED ne 'N' then

i. **GP = .**

ii. **UNRES** = 1

b. If MAFID = 'XXXXXXXXX' then set GP = . and UNRES = 1. Obtain MAFID from GQCI team.

Commented [JEZ(F2]: Issue #1 with all reported pop in one dorm.

c. Otherwise,

i. **GP** = GQ_INITIAL_POP

ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

a. Calculate GP/GQ_SIZE_EXP_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.
   We will create 3 ratios comparing GP to GQ_SIZE_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):
   1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
   2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
   3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
   4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
   5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
   6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

ii. Calculate Bounds.
   For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_EXP_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGA in (' ','R'). Assign these values as **EXP_PERS_10** and **EXP_PERS_90** respectively.
   For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_EXP_PERS_CNT.
   1. Assign **EXP_PERS_TRUNC** = GQ_SIZE_EXP_PERS_CNT
   2. If GQ_SIZE_EXP_PERS_CNT > EXP_PERS_90 then set **EXP_PERS_TRUNC** = EXP_PERS_90
   3. If GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_EXP_PERS_CNT < EXP_PERS_10 then set **EXP_PERS_TRUNC** = EXP_PERS_10.

iii. Assign values. For each MAFID, calculate the following values:
   1. **IMP_RAT_EXP** = CEIL (EXP_PERS_TRUNC*EXPRATIO)
   2. **IMP_RAT_EXP_GQ** = CEIL (EXP_PERS_TRUNC*EXPRATIO_GQ)
   3. **IMP_RAT_EXP_GQ_ST** = CEIL (EXP_PERS_TRUNC*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.
   We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):
   1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
   2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
   3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
   4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID

5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_MAX_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGB in (' ','R'). Assign these values as **MAX_PERS_10** and **MAX_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_MAX_PERS_CNT.

1. Assign **MAX_PERS_TRUNC** = GQ_SIZE_MAX_PERS_CNT
2. If GQ_SIZE_MAX_PERS_CNT > MAX_PERS_90 then set **MAX_PERS_TRUNC** = MAX_PERS_90
3. If GQ_SIZE_MAX_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT < MAX_PERS_10 then set **MAX_PERS_TRUNC** = MAX_PERS_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_MAX** = CEIL (MAX_PERS_TRUNC*MAXRATIO)
2. **IMP_RAT_MAX_GQ** = CEIL (MAX_PERS_TRUNC*MAXRATIO_GQ)
3. **IMPRAT_MAX_GQ_ST** = CEIL (MAX_PERS_TRUNC*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):

1. Sum the GP and GQCURRSIZE value **for the nation.**
2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRSIZE for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGC in (' ','R'). Assign these values as **CURRSIZE_10** and **CURRSIZE_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRSIZE.

1. Assign **CURRSIZE_TRUNC** = GQCURRSIZE
2. If GQCURRSIZE > CURRSIZE_90 then set **CURRSIZE_TRUNC** = CURRSIZE_90

3. If GQCURRSIZE > 0 and GQCURRSIZE < CURRSIZE_10 then set **CURRSIZE_TRUNC** = CURRSIZE_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURR** = CEIL (CURRSIZE_TRUNC*CURRSIZERATIO)
2. **IMP_RAT_CURR_GQ** = CEIL (CURRSIZE_TRUNC*CURRSIZERATIO_GQ)
3. **IMP_RAT_CURR_GQ_ST** = CEIL (CURRSIZE_TRUNC*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRMAXPOP Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRMAXPOP for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGD in (' ','R'). Assign these values as **CURRMAX_10** and **CURRMAX_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRMAXPOP.

1. Assign **CURRMAX_TRUNC** = GQCURRMAXPOP
2. If GQCURRMAXPOP > CURRMAX_90 then set **CURRMAX_TRUNC** = CURRMAX_90
3. If GQCURRMAXPOP > 0 and GQCURRMAXPOP < CURRMAX_10 then set **CURRMAX_TRUNC** = CURRMAX_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (CURRMAX_TRUNC*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (CURRMAX_TRUNC*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (CURRMAX_TRUNC*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):

- i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**
- ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**
- iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**
    1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
    2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
    3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**
- iv. Assign values. For each MAFID, calculate the following values:
    1. **IMP_MEDGP_GQ_ST** = CEIL(MEDGP_GQ_ST)
    2. **IMP_MEDGP_GQ** = CEIL(MEDGP_GQ)
    3. **IMP_MEDGP** = CEIL(MEDGP)

C. CES method: impute using a hybrid of the ratio imputes created in the previous step, a percentile method based on Greek/non-Greek status, and a facility-level residual allocation method.

- a. Ingest the file referred to as **MAFID_FRAT_SORO**
    - i. On this file **FLAG_GREEK_LETTER**=1 indicates that GQ has been identified as a fraternity or sorority house. Otherwise **FLAG_GREEK_LETTER**=0.
- b. Ingest the file referred to as **UNITID_MAFID_LINKS**.
    - i. When reading in **UNITID_MAFID_LINKS,** keep only the variables **MAFID, UNITID, MATCH_STEP_NUM**, and **ROOMCAP.**
    - ii. Note: for records with **MATCH_STEP_NUM**=-1, **UNITID** will be missing.
    - iii. Note: for records with the same value of UNITID, ROOMCAP will be the same.
- c. Merge **MAFID_FRAT_SORO** and **UNITID_MAFID_LINKS** to ***GQ_MAFID***, merging on MAFID, and keeping only records that are in ***GQ_MAFID.***
    - i. Note: For records that match, this should be a 1-to-1 match (MAFID should be unique in each of the 3 datasets).
    - ii. Note: only records with GQCURTYP=501 in ***GQ_MAFID*** should match to either of the other 2 datasets.
- d. Select the subset of the merged dataset from the previous step with GQCURTYP=501.
    - i. NOTE: In this spec we will refer to this subset of the data as **GQ_COUNTS_ROOMCAP_GREEK**. This is only an intermediate dataset, which will be merged back to the **GQ_MAFID** dataset at the end of this section of the spec (section 5.D).
- e. Using GQ_COUNTS_ROOM_CAP_GREEK and the ratio impute variables created in section 4.A, create a temporary impute variable IMP_GP_TEMP using the hierarchy shown in the following table. If IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP_TEMP= IMP_RAT_EXP_GQ_ST and set ALREADY_IMPUTED=1. If IMP_RAT_EXP_GQ_ST is missing and IMP_RAT_EXP_GQ is not missing, assign IMP_GP_TEMP= IMP_RAT_EXP_GQ and set ALREADY_IMPUTED=1. Continue through the table until all the variables in the table have been exhausted. For any remaining

MAFIDs for which a value has not been assigned to IMP_GP_TEMP, set ALREADY IMPUTED=0;

IMP_GP_TEMP assignment hierarchy
IMP RAT EXP GQ ST
IMP RAT EXP GQ
IMP RAT MAX GQ ST
IMP RAT MAX GQ
IMP RAT CURR GQ ST
IMP RAT CURR GQ
IMP RAT CURRMAX GQ ST
IMP RAT CURRMAX GQ

f. Using only MAFIDs in **GQ_COUNTS_ROOMCAP_GREEK** with UNRES = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’,’R’), create 3 GP median variables and 3 GP maximum variables:
    i. For each UNITID-FLAG_GREEK_LETTER combination:
        1. Calculate the median value of GP. Call this **P50_GP_UNIT_BY_GRK**
        2. Calculate the maximum value of GP. Call this **MAX_GP_UNIT_BY_GRK.**
        3. Merge the P50_GP_UNIT_BY_GRK and MAX_GP_UNIT_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on UNITID and FLAG_GREEK_LETTER.
        4. Note, these values will be missing if there are not enough observations for the UNITID-FLAG_GREEK_LETTER combination.
    ii. For each BCUSTATEFP-FLAG_GREEK_LETTER combination:
        1. Calculate the median value of GP. Call this **P50_GP_ST_BY_GRK**.
        2. Calculate the maximum value of GP. Call this **MAX_GP_ST_BY_GRK**.
        3. Merge P50_GP_ST_BY_GRK and MAX_GP_ST_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on BCUSTATEFP-FLAG_GREEK_LETTER combinations.
        4. Note, these values will be missing if there are not enough observations for the BCUSTATEFP-FLAG_GREEK_LETTER combination.
    iii. For each value of FLAG_GREEK_LETTER:
        1. Calculate the median value of GP. Call this **P50_GP_BY_GRK.**
        2. Calculate the maximum value of GP. Call this **MAX_GP_BY_GRK**.
        3. Merge P50_GP_BY_GRK and MAX_BP_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on FLAG_GREEK_LETTER.

g. For MAFIDs for which UNRES=1, FLAG_GREEK_LETTER=1, and ALREADY_IMPUTED=0, assign median Greek imputes to IMP_GP_TEMP and create up to 3 new impute variables using the following hierarchy:
    i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:
        1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
        2. Set ALREADY_IMPUTED=1
        3. Assign **MEDGP_GRK_UNIT**= IMP_GP_TEMP
    ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:
        1. assign IMP_GP_TEMP= P50_GP_ST_BY_GRK

            2. set ALREADY_IMPUTED=1
            3. Assign **MEDGP_GRK_ST**= IMP_GP_TEMP
        iii. Otherwise:
            1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
            2. Set ALREADY_IMPUTED=1
            3. Assign **MEDGP_GRK**=IMP_GP_TEMP

h. Using **GQ_COUNTS_ROOMCAP_GREEK**, by UNITID, create unit-level sum variables (where a unit corresponds to a single UNITID, which corresponds to a single a university or college)
    i. Create unit-level sums (i.e., by UNITID) of GQCURRMAXPOP using only observations where flagD in (‘’,’R’). Note: these are the “good” values of GQCURRMAXPOP. Note that for this sum, we don’t care what the value of GP is, even it is a true 0. We are just trying to come up with a maximum number of people that these GQs *could* house, so that we can subtract the sum from the college-level IPEDS ROOMCAP variable. For reference later in the spec, call this sum **UNIT_MAXPOP_SUM**.

    ii. Using only the GQs with unres=0 and flagA not in ('I',’S’) and flagB not in ('I',’S’) and flagC not in ('I',’S’) and flagD = 'M' and GQCURRMAXPOP=., by UNITID, create unit-level sums of GP. Call this sum **UNIT_2020POP_SUM**.
    iii. Using only the GQs with (unres=1 or flagA = 'I' or flagB='I' or flagC='I' or flagD='I') and already_imputed=1 and GQCURRMAXPOP=., by UNITID, create unit-level sums of IMP_GP_TEMP. Call this **UNIT_POP_IMPUTED_SUM**.
    iv. Create **UNIT_CAP_SUM** = the unit-level sum of UNIT_MAXPOP_SUM, UNIT_2020POP_SUM, and UNIT_POP_IMPUTED_SUM

i. For each MAFID, calculate UNIT_RESIDUAL = ROOMCAP – UNIT_CAP_SUM (this will be the same value for MAFIDs with the same UNITID)

j. For each MAFID with UNIT_RESIDUAL<=0, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP, and create 3 new (non-Greek) median impute variables using the following hierarchy:
    i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:
        1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
        2. Set ALREADY_IMPUTED=1
        3. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP
    ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:
        1. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
        2. Set ALREADY_IMPUTED=1
        3. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP
    iii. Otherwise:
        1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
        2. Set ALREADY_IMPUTED=1
        3. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

k. For each (non-missing) UNITID with UNIT_RESIDUAL>0, count the MAFIDs associated with that UNITID that have UNRES=1 and ALREADY_IMPUTED=0. Call this count UNIT_RESID_GQ_COUNT.

l. For MAFIDs with UNIT_RESIDUAL>0, UNIT_RESID_GQ_COUNT=1, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy:
   - i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK, then assign values to IMP_GP_TEMP using the following sub-hierarchy:
     - 1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:
       - a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
       - b. Set ALREADY_IMPUTED=1
       - c. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP
     - 2. Otherwise (i.e., if P50_GP_UNIT_BY_GRK<=0 or missing), if MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing, then:
       - a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
       - b. Set ALREADY_IMPUTED=1
       - c. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP
     - 3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:
       - a. Assign IMP_GP_TEMP= P50_GP_BY_GRK
       - b. Set ALREADY_IMPUTED=1
       - c. Assign **MEDGP_nonGRK**=IMP_GP_TEMP
   - ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL < MAX_GP_UNIT_BY_GRK, then assign values as follows:
     - 1. Assign IMP_GP_TEMP=UNIT_RESIDUAL
     - 2. Set ALREADY_IMPUTED=1
     - 3. Assign **IMP_RESID_1GQ**=IMP_GP_TEMP

m. For MAFIDs with UNIT_RESIDUAL>0, UNIT_RESID_GQ_COUNT>1, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy. (NOTE: steps i.1-i.3 are the same as steps i.1-i.3 in step l above):
   - i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK, then assign values to IMP_GP_TEMP using the following sub-hierarchy:
     - 1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:
       - a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
       - b. Set ALREADY_IMPUTED=1
       - c. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP
     - 2. Otherwise (i.e., if P50_GP_UNIT_BY_GRK<=0 or missing), if MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing, then:
       - a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
       - b. Set ALREADY_IMPUTED=1
       - c. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP
     - 3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:
       - a. Assign IMP_GP_TEMP= P50_GP_BY_GRK
       - b. Set ALREADY_IMPUTED=1
       - c. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

   ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL < MAX_GP_UNIT_BY_GRK, then assign values as follows:
      1. Assign IMP_GP_TEMP=UNIT_RESIDUAL/UNIT_RESID_GQ_COUNT
      2. Set ALREADY_IMPUTED=1
      3. Assign **IMP_RESID_NGQ**=IMP_GP_TEMP

n. Do a cross-tabulation of the variables UNRES and ALREADY_IMPUTED. If ALREADY_IMPUTED is always 1 when UNRES=1, then imputations have been calculated for all MAFIDS with GQCURTYP 501.

o. Keep the variables **MEDGP_GRK_UNIT, MEDGP_GRK_ST, MEDGP_GRK, MEDGP_nonGRK_UNIT, MEDGP_nonGRK_ST, MEDGP_nonGRK, IMP_RESID_1GQ**, and **IMP_RESID_NGQ.** Drop all other variables created in this section

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table hierarchically as follows, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. If IMP_RAT_EXP_GQ_ST is missing, if IMP_RAT_EXP_GQ is not missing, assign IMP_GP = IMP_RAT_EXP_GQ and assign IMP_FLAG = 102. Continue on through the table until all MAFIDs with UNRES = 1 have a value for IMP_GP and IMP_FLAG.

IMP GP	**IMP FLAG**
IMP RAT EXP GQ ST	101
IMP RAT EXP GQ	102
IMP RAT EXP	103
IMP RAT MAX GQ ST	104
IMP RAT MAX GQ	105
IMP RAT MAX	106
IMP RAT CURR GQ ST	107
IMP RAT CURR GQ	108
IMP RAT CURR	109
IMP RAT CURRMAX GQ ST	110
IMP RAT CURRMAX GQ	111
IMP RAT CURRMAX	112
MEDGP GRK UNIT	301
MEDGP GRK ST	302
MEDGP GRK	303
MEDGP nonGRK UNIT	304
MEDGP nonGRK ST	305
MEDGP nonGRK	306
IMP RESID 1GQ	307
IMP RESID NGQ	308
IMP MEDGP GQ ST	401
IMP MEDGP GQ	402
IMP MEDGP	403

Section 6: Create Output Files

Output the following variables from GQMAFID:

Commented [JEZ(F3]: Ryan's recent files don't have geography on them...

MAFID	ACOCE	BCUCOUNTYFP
BCUSTATEFP	FACTLNAME	GQ SIZE EXP PERS CNT
GQ SIZE MAX PERS CNT	GQCONTACT	GQCURRMAXPOP
GQCURRSIZE	GQNAME	GQTYPCUR
GQ INITIAL STATUS	GQ INITIAL UNRES	GQ INITIAL POP
IMPUTE NEEDED	FLAGA	FLAGB
FLAGC	FLAGD	GP
UNRES	IMP GP	IMP FLAG
EXPRATIO	EXPRATIO GQ	EXPRATIO GQ ST
EXP PERS 10	EXP PERS 90	EXP PERS TRUNC
IMP RAT EXP	IMP RAT EXP GQ	IMP RAT EXP GQ ST
MAXRATIO	MAXRATIO GQ	MAXRATIO GQ ST
MAX PERS 10	MAX PERS 90	MAX PERS TRUNC
IMP RAT MAX	IMP RAT MAX GQ	IMP RAT MAX GQ ST
CURRRATIO	CURRRATIO GQ	CURRATIO GQ ST
CURRSIZE 10	CURRSIZE 90	CURRSIZE TRUNC
IMP RAT CURR	IMP RAT CURR GQ	IMP RAT CURR GQ ST
CURRMAXRATIO	CURRMAXRATIO GQ	CURRMAXRATIO GQ ST
CURRMAX 10	CURRMAX 90	CURRMAX TRUNC
IMP RAT CURRMAX	IMP RAT CURRMAX GQ	IMP RAT CURRMAX GQ ST
MEDGP	MEDGP GQ	MEDGP GQ ST
IMP MEDGP	IMP MEDGP GQ	IMP MEDGP GQ ST
MEDGP GRK UNIT	MEDGP GRK ST	MEDGP GRK
MEDGP nonGRK UNIT	MEDGP nonGRK ST	MEDGP nonGRK
IMP RESID 1GQ	IMP RESID NGQ	

Name this file gq_mafid_dssd_out_validation.sas7bdat

Output the following variables from GQMAFID:

MAFID	ACOCE	BCUCOUNTYFP
BCUSTATEFP	FACTLNAME	GQ SIZE EXP PERS CNT
GQ SIZE MAX PERS CNT	GQCONTACT	GQCURRMAXPOP
GQCURRSIZE	GQNAME	GQTYPCUR
GQ INITIAL STATUS	GQ INITIAL UNRES	GQ INITIAL POP
IMPUTE NEEDED	FLAGA	FLAGB
FLAGC	FLAGD	GP
UNRES	IMP GP	IMP FLAG
CALL STATUS	GEO POP COUNT	

Name this file gq_mafid_dssd_out_pop.sas7bdat. See POP data dictionary.

Survey Methodology, June 1986
Vol. 12, No. 1, pp. 73–83
Statistics Canada

Statistical Editing and Imputation for Periodic Business Surveys

M.A. HIDIROGLOU and J.-M. BERTHELOT[1]

ABSTRACT

For periodic business surveys which are conducted on a monthly, quarterly or annual basis, the data for responding units must be edited and the data for non-responding units must be imputed. This paper reports on methods which can be used for editing and imputing data. The editing is comprised of consistency and statistical edits. The imputation is done for both total non-response and partial non-response.

KEY WORDS: Periodic survey; Statistical editing; Total/partial non-response; Imputation.

1. INTRODUCTION

Data are routinely collected by large organizations such as Statistics Canada based on properly designed sample surveys. If such data are collected on a periodic basis from the same sampling unit, there are several possibilities which will occur with respect to the data consistency (quality) over a given time period. The sampling unit may report the data faithfully with no dramatic departure in continuity ("smoothness") as time progresses. The data may be reported faithfully, with questionable jumps between two time periods. The sampling unit may not report all the requested data items: this is known as partial non-response. The sampling unit may report data sporadically with breaks of total non-response for some periods. These can occur simultaneously in a periodic survey which collects required data from a large number of sampling units.

The problems which will be addressed in this article are the editing and imputation of data for sampling units that are contacted on a periodic basis by a surveying organization. The methods discussed are general for data of a multivariate nature composed of both quantitative and qualitative variables. The editing will include consistency and statistical edits.

For quantitative data, consistency edits ensure that linear combination of the data fields within a given time period satisfy given requirements. For qualitative data, consistency edits ensure that variables correspond to well defined values.

Statistical edits are used to isolate sampling units which may report some of their quantitative data fields in an inconsistent manner either from time period to time period or within a specific time period. Units with unusually high or low values will be termed "outliers". The identification of "outliers" is extremely important in an ongoing survey for two reasons. First, they influence statistics of the data set which may be for instance totals. This point has been studied by Hidiroglou and Srinath (1981). Second, the imputation of quantitative data for non-response units for periodic business surveys is usually based on trends or means: the removal of outlier units from the computation of these trends or means, will produce statistics that are not contaminated with there observations. For units which have partial non-response, data must be imputed for the missing fields.

For large data sets, where timely release of the summary information is crucial, the editing and the imputation of data should be automatic and computer handled given some well specified rules. This is in agreement with Gentleman and Wilk (1975), and Fellegi and Holt (1976).

[1] M.A. Hidiroglou and J.-M. Berthelot, Business Survey Methods Division, 11th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

2. EDITING PERIODIC DATA

2.0 Consistency Edits

For a given unit i and time period t, let $\underset{\sim}{x}_i(t)$ represent the vector of data which is to be collected. The vector $\underset{\sim}{x}_i(t)$ may be decomposed into a series of elementary vectors for which independent editing and imputation are required.

That is, $$\underset{\sim}{x}_i(t) = (\underset{\sim}{x}_i^{(1)}(t), \ldots, \underset{\sim}{x}_i^{(P)}(t))$$

where $$\underset{\sim}{x}_i^{(p)}(t) = (x_{il}^{(p)}(t), \ldots, x_{ik_p}^{(p)}(t))$$

for $$i=1, \ldots, n;\ p=1, \ldots, P;\ t=1, \ldots, T$$

and k_p is the number of variables in the p:th elementary vector.

For each elementary vector $\underset{\sim}{x}_i^{(p)}(t)$, the consistency edits may be represented as

$$\underset{\sim}{A}^{(p)}(\underset{\sim}{x}_i^{(p)}(t))' \leq (\underset{\sim}{c}^{(p)})'$$

where $\underset{\sim}{A}^{(p)}$ is a ℓ_p by k_p matrix representing the rules that the elements of the elementary vector $\underset{\sim}{x}_i^{(p)}(t)$ must obey, and $\underset{\sim}{c}^{(p)}$ is a 1 by ℓ_p vector which represents the constraints. This formulation allows one to define consistency edits for both qualitative and quantitative variables. For qualitative variables, the consistency edits could be used to check if the variables correspond to well-defined values. For quantitative variables, the consistency edits can check if certain variables are not larger (or smaller) than other variables or that a linear combination is equal to (or greater than or less than) a given variable.

2.1 Statistical Edits

Given that data are reported periodically, the problem is to isolate outlying observations within the time series. In the present context, an outlying observation i, will be defined as one whose trend for the current period to a previous period, for given variables of the element vector $\underset{\sim}{x}_i(t)$, differs significantly from the corresponding overall trend of other observations belonging to the same subset of the population. Statistical edits can also be applied within a time period, by comparing the ratios of two correlated variables amongst themselves, within a given subset of the population. In this article, the statistical edit will only be discussed in terms of the trend between time periods. Similar, somewhat imprecise but working definitions of outliers have also been given by other authors, for example:

GRUBBS (1969) says that "An outlying observation, or outlier, is one that appears to deviate markedly from the other members of the sample in which it occurs."

GUMBEL (1960) says: "The outliers are values which seem either too large or too small as compared to the rest of the observations."

KENDALL and BUCKLAND (1957, p. 209), write: "In a sample of n observations it is possible for a limited number to be so far separated in value from the remainder that they give rise to the question whether they are from a different population, or that the sampling technique is at fault. Such values are called outliers. Tests are available to ascertain whether they can be accepted as homogeneous with the rest of the sample."

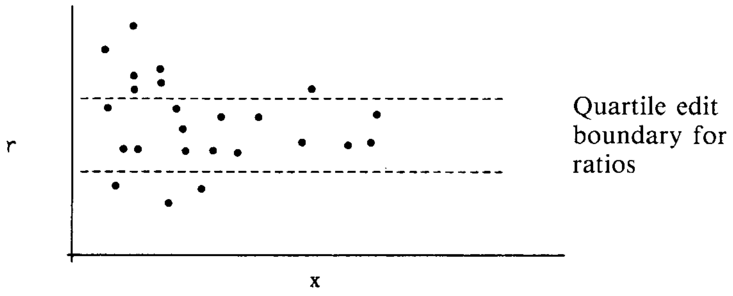
2.1.1 Review of Some Methods Currently Used

Methods for detecting outliers have been proposed by Dixon (1953), Grubbs (1969), Tietgen and Moore (1972), and Prescott (1978) to mention a few. Most of the test procedures for outlier detection proposed by these authors consider the problem as one of hypothesis testing. In the simplest cases, the null hypothesis is that the sample comes from a normal distribution with unspecified mean and variance, while the alternative hypothesis is that one or more of the observations come from a different distribution. Percentage points of a test statistic may be determined under the null hypothesis and compared with computed values of the test statistic in particular applications. Applying these methods to periodic data from large surveys presents problems for the following reasons. First, the assumption of normality of trends from one period to another may not hold. Second, these traditional methods require the existence of tables for determining critical values which define rejection regions. The method which we will propose in Section 2.1.2 does not have the above mentioned disadvantages. It can be easily implemented on the computer, does not require the assumption of normality, and does not make use of tables.

In our specific context, and given elements of the vectors $\underset{\sim}{x}_i(t)$ and $\underset{\sim}{x}_i(t + 1)$, denote as $x_i(t)$ and $x_i(t + 1)$ the responses for two consecutive periods for a given unit, where $i=1, ..., n$. Denote as r_i the ratio of current period data to previous period data. One method which is known as the range edit, is to simply define fixed upper and lower bounds based on experience for comparison purposes. Ratios found outside these bounds are declared as outliers. A major drawback with this method is that the definition of outlier is too subjective and does not make use of the distribution of the ratios.

A method that attempts to make use of the distribution of the ratios is the Chebychev inequality edit. This edit is constructed by computing the lower bound as $\bar{r} - ks_r$ and the upper bound as $\bar{r} + ks_r$ where $\bar{r} = \Sigma_{i=1}^n r_i/n$ and $s_r^2 = \Sigma_{i=1}^n (r_i-\bar{r})^2/(n-1)$. This edit has two main drawbacks. First, the choice of k is subjective and can result in having an edit that cannot detect any outliers. This last point has been demonstrated by Wilkinson (1982). Second, "large" outliers may hide "smaller" outliers. This effect is known as the masking effect.

An improvement to this method has been the use of quartiles and interquartile distances rather than the use of mean and standard error to come up with the upper and lower bounds. In this case, the edit is constructed by computing the lower bound as $r_M - k D_{r_{Q1}}$ and the upper bound as $r_M + k D_{r_{Q3}}$ where r_M is the median of the ratios, $D_{r_{Q1}}$is the distance between the first quartile and the median, and $D_{r_{Q3}}$ is the distance between the third quartile and the median. Since the quartiles are not affected by the tails of the distribution, it greatly alleviates the masking effect problem. However, this method has two drawbacks. First, in some very specific circumstances, it is possible that the outliers on the left tail of the distribution are undetectable. Second this method does not take into account the fact that in most of the periodic business surveys, the variability of ratios for small businesses is larger than the variability of ratios for large businesses (Sugavanam 1983). This fact is expressed by the following graph:



This drawback has the effect of identifying too many small units as outliers and not enough large units. This effect will be referred to as the "size masking effect".

2.1.2 Proposed Procedure

For two occasions t and $t + 1$, the overall trend for the data pair given by

$$(x_i(t), x_i(t + 1)), i = 1, ..., n$$

is

$$R = \sum_{i=1}^{n} x_i(t + 1) / \sum_{i=1}^{n} x_i(t).$$

Now, R may be expressed as

$$R = \sum_{i=1}^{n} I_i r_i$$

where

$$I_i = x_i(t) / \sum_{i=1}^{n} x_i(t)$$

and

$$r_i = x_i(t + 1)/x_i(t).$$

I_i is a measure of the relative importance of the *i:th* unit amongst the n units at time t. The individual trends r_i must be transformed in order to ensure that outliers are detected at both tails of the distribution. This transformation is:

$$s_i = \begin{cases} 1 - r_M/r_i, \text{ if } 0 < r_i < r_M \\ r_i/r_M - 1, \text{ if } r_i \geq r_M \end{cases}$$

where r_M is the median of the ratios.

In order to bring in the magnitute of the data, the following transformation is required (Berthelot 1983):

$$E_i = s_i \{\text{Max } (x_i(t), x_i(t + 1))\}^U$$

where $0 \leq U \leq 1$. The E_i's will be referred to as effects and the exponent U in the transformation provides a control on the importance associated with the magnitude of the data. This transformation allows us to place more importance on a small change associated with a "large" unit as opposed to a large change associated with a "small" unit. The values of the median and quartiles as used by Sande (1981) will be applied to the transformed, E_i's, in order to detect potential outliers. Denoting as E_{Q1}, E_M and E_{Q3} as the first quartile, the median and the third quartile respectively, define the following two deviations:

$$d_{Q1} = \text{Max } (E_M - E_{Q1}, |AE_M|),$$

$$d_{Q3} = \text{Max } (E_{Q3} - E_M, |AE_M|).$$

Outliers will be defined as all those units whose associated effect E_i lies outside the interval $(E_M - Cd_{Q1}, E_M + Cd_{Q3})$. The purpose of the AE_M term is to avoid difficulties which arise when $E_M - E_{Q1}$ or $E_{Q3} - E_M$ are very small. That is, the problem which may arise when the effects E_i are clustered around a single value with one or two modest deviations may produce false outliers. The parameter C controls the width of the acceptance interval. The parameter U controls the shape of the curve defining upper and lower boundaries. The effect of increasing U is to attach more importance with fluctuations associated with the larger observations. A value of 0.05 is suggested for A as it has proved to be adequate in practice.

2.1.3 Treatment For Outliers

Once units have been identified as possible outliers, they are flagged as such and brought to the attention of the survey takers. A decision must then be taken on how these abnormal observations are treated. Their existence may have arisen as a result of several factors. These factors include measurement error, incorrect interpretation of the questionnaire by the responding unit, or intrinsic variability of the population being surveyed. For units which have measurement error due to incorrect transcription of the data or incorrect responses, a simple follow-up will clear up the majority of these errors. For units which display intrinsic variability as a result of rapid growth, the reported values are correct but dominate too much the resulting summary tables. For those units, techniques, which reduce the sampling weight as suggested by Hidiroglou and Srinath (1981) or change the values themselves as suggested by Ernst (1980), must be used in order to accomodate (minimize) the effect of outlying observations. For units having unrepresentative data which cannot be verified, their data must be substituted with other data based on imputation techniques. The different kinds of corrective actions taken on outlying units must be flagged as well.

3. IMPUTING PERIODIC DATA

The information collected by periodic business surveys, such as sales and employment are collected via samples using mail questionnaires or telephone interviews. Non-responding units are followed up as much as possible within allotted budgets in order to improve the response rates. The follow-up is usually done by mail in the case of the smaller to medium sizes non-responding companies and by telephone for the larger or dominating companies. Although following up delinquent companies improves response rates for a given reference period, there will be nevertheless, a group of non-responding companies which may be classified into either hard-core or late respondents. Hard-core non-respondents are units which require a great deal of persuation to respond, if at all. Late respondents are units which respond late with respect to the survey's reference period either because they do not mail back their questionnaire on time or because they need to be prompted by a follow-up questionnaire. The non-responding units must therefore be imputed in order to make up for their contribution to the particular estimator being used by the survey. In the case of Monthly Business Surveys, such as the Monthly Retail Trade Survey, totals (e.g., sales) are being estimated. Imputation procedures can also be used to generate values for units declared as outliers. These imputed values can be used in lieu of these outlying observations, if no valid explanation can be provided for their presence.

The units with no response whatsoever, will be termed as total non-respondents and those with some, but not all, required data items, will be termed partial non-respondents. Desirable features of an imputation system should include the following properties (Berthelot and Hidiroglou 1982):

- it must automatically determine the most reasonable imputation procedure possible under the existing circumstances,
- the imputation cell, the level at which the computation of trends and means (medians) is performed, will usually correspond to the finest level of stratification of the sample,
- a minimum number of units must participate in the computation of trends or means (medians), otherwise, the imputation cells are automatically collapsed (using a pre-determined pattern), until the minimum requirement has been satisfied,
- it will recognize through the use of status codes that there are units which must not be imputed. These include seasonal units during the period that they are not operating, units temporarily out of business, or units which are no longer active,
- births which have no previous business history will have their data imputed using the means (medians) of similar responding births,
- units will be re-imputed for a number of periods previous to the current period: this is done in order to improve the strength of the imputations if the previous periods have been updated with data,
- backward imputations will be applied to units which have been continuously imputed using a forward imputation procedure as soon as a good response is obtained for a given period,
- imputation status codes will be associated with imputed units in order to provide a history of the procedure used for imputation,
- the ranking for imputing non-responding units is as follows: trends (monthly, quarterly, annual), means (medians) with the most recent trends being given priority. For instance, in the case of a monthly system, monthly trends are used for units which have data (response or imputed) in the month prior to the one to be imputed. Annual trends are used mostly for units which are seasonal and which fail to provide a response as they emerge from their out of season period and for which a last year value existed for the month to be imputed. Imputations based on the trends are obtained by multiplying the trends by the unit's last month or last year value. In the event that trends cannot be applied, the mean (median) of the cell is used as an imputation.

In order to formalize the preceding paragraphs in a mathematical fashion, let the number of units which are expected to respond for a given cell and given month be n. Let the number of non-respondents with total non-response be n_3, the number of respondents with total response be n_1 and the number of respondents with partial response be n_2. It is assumed that the sample design is stratified with the sampling being simple random without replacement. Let the size for the follow-up sample of the non-respondents be m_3 ($2 \leq m_3 \leq n_3$, with m_3 having been selected from n_3 according to a randomized mechanism). Note that $n_4 = n - \Sigma_{i=1}^{3} n_i$ units are not expected to provide any response to the survey process for a number of possible reasons. At a time t, they may be out of season, inactive, dead, or out of scope to the survey. For these units, the system will automatically associate zero values for all relevant fields in the given period.

The imputation process will then be done in several different ways according to the type of non-response.

3.0 Total Non-Response

The imputation process for the total non-respondents will first be discussed. Bearing in mind that either the whole vector $\underset{\sim}{x}_i(t)$ or that some of its elementary vectors as given in

Section 2.0 must be totally imputed, denote as $(x_{i1}(t), \ldots, x_{ip}(t))$ one of the elementary vector within $\underset{\sim}{x}_i(t)$ where the editing and imputation process is independent from other elementary vectors within $\underset{\sim}{x}_i(t)$. Assuming that

$$x_{ip}(t) \geq \sum_{j=1}^{p-1} x_{ij}(t),$$

(which implies that the sum of the first $p-1$ data elements of the elementary vectors are smaller than the *p:th* datum element, the total) $x_{ip}(t)$ will first be imputed as

$$I_{ip}^{(1)}(t) = \sum_{k=1}^{6} [z_{ip}^{(k)}(t)\ \delta_i^{(k)}]$$

where $\delta_i^{(k)}$ refers to the procedure used for imputation and $z_{ip}^{(k)}$is the associated imputed value. One of the six $\delta_i^{(k)}$ values will be one and the other five must be zero $(\Sigma_{k=1}^{6}$ $\delta_i^{(k)} = 1)$. The imputed $z_{ip}^{(k)}(t)$ values will be as follows:

$$z_{ip}^{(1)}(t) = [\sum_{r\epsilon s_1} w_r\, x_{rp}(t) / \sum_{r\epsilon s_1} w_r\, x_{rp}(t-1)]\ x_{ip}(t-1),$$

$$z_{ip}^{(2)}(t) = [\sum_{r\epsilon s_2} w_r\, x_{rp}(t) / \sum_{r\epsilon s_2} w_r\, x_{rp}(t-Q)]\ x_{ip}(t-Q),$$

$$z_{ip}^{(3)}(t) = [\sum_{r\epsilon s_3} w_r\, x_{rp}(t) / \sum_{r\epsilon s_3} w_r\, x_{rp}(t-1)]\ x_{ip}(t-1),$$

$$z_{ip}^{(4)}(t) = [\sum_{r\epsilon s_4} w_r\, x_{rp}(t) / \sum_{r\epsilon s_4} w_r\, x_{rp}(t-Q)]\ x_{ip}(t-Q),$$

$$z_{ip}^{(5)}(t) = [\sum_{r\epsilon s_5} w_r\, x_{rp}(t) / \sum_{r\epsilon s_5} w_r],$$

$$z_{ip}^{(6)}(t) = [\sum_{r\epsilon s_6} w_r\, x_{rp}(t) / \sum_{r\epsilon s_6} w_r],$$

w_r = inverse selection probability of unit r for the given cell. The subsets s_i $(i=1, \ldots, 6)$, will be determined by selecting the units which have provided a response for the *p:th* variable at time t and which have passed the edits. The conditions for each subset is

s_1 = all units which have provided edited responses between times t and $t-1$,

s_2 = all units which have provided edited responses between times t and $t-Q$,

s_3 = units in the follow-up subsample which have provided edited responses between times t and $t-1$,

s_4 = units in the follow-up subsample which have provided edited responses between times t and $t-Q$,

s_5 = all units which have provided edited responses at time t,

s_6 = units in the follow-up subsample which have provided edited responses at time t.

The choice of the imputation procedure will be governed by the following considerations.

(i) Procedures 1 (or 2) will be used if there is a response or imputed value at time $t-1$ (or $t-Q$) and that it is believed that the trends for the non-respondents is the same as the one for the respondents, within the given cell,

(ii) Procedures 3 (or 4) will be used if there is a response or imputed value at time $t-1$ (or $t-Q$) and that it is believed that the trends for the non-respondents differs from the one for the respondents within the given cell.

(iii) Procedure 5 will be used if there is no response at either times $t-1$ or $t-Q$ and that is believed that the mean of the non-respondents is equal to the mean of the respondents within the given cell,

(iv) Finally, procedure 6 will be used if there is no response at either times $t-1$ or $t-Q$ and that it is believed that the means of the respondents and non-respondents are different.

The choices between the different procedures can be made using decision tables which determine the conditions and, given the condition, choose the best imputation procedure according to pre-determined rules. Once that $x_{ip}(t)$ has been imputed for an elementary vector, its remaining components can be imputed using the procedures for partial non-response.

3.1 Partial Non-Response

For an elementary vector $(x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))$ which is part of $\underset{\sim}{x}_i(t)$, let δ_{ij} be the indicator variable which is equal to 1 if $x_{ij}(t)$ is present and zero otherwise at time t. Some additional notation is introduced at this point in order to ease the development. To this end, define

$$s_{i,R}(t-1) = \sum_{j=1}^{p-1} \delta_{ij}\, x_{ij}(t-1)$$

= the sum of responses at time $t-1$, for which there is a response at time t

$$s_{i,NR}(t-1) = \sum_{j=1}^{p-1} (1-\delta_{ij})\, x_{ij}(t-1)$$

= the sum of responses at time $t-1$, for which there is no response at time t,

$$s_{i,R}(t) = \sum_{j=1}^{p-1} \delta_{ij}\, x_{ij}(t).$$

The partial imputation will be based on the assumptions that $x_{ip}(t) \geq \Sigma_{j=1}^{p-1} x_{ij}(t)$ and that the distribution of the elements within $\underset{\sim}{x}_i(t)$ is similar to the distribution of the elements within $\underset{\sim}{x}_i(t-1)$. Two separate cases will be discussed.

Case 1: Parts of the elementary vector missing and $x_{ip}(t)$ present

Two subcases are possible: $x_{ip}(t) = \sum_{j=1}^{p-1} x_{ij}(t)$ or $x_{ip}(t) > \sum_{j=1}^{p-1} x_{ij}(t)$.

(i) $x_{ip}(t) = \sum_{j=1}^{p-1} x_{ij}(t)$

If all the elements of $\underset{\sim}{x}_i(t)$ excluding $x_{ip}(t)$ are missing, that is $\Sigma_{j=1}^{p-1} \delta_{ij} = 0$, then we must have that $s_{i,NR}(t) = x_{ip}(t)$. If some of the elements of $\underset{\sim}{x}_i(t)$ excluding $x_{ip}(t)$ are missing, that is $\Sigma_{j=1}^{p-1} \delta_{ij} > 0$, then $s_{i,NR}(t) = x_{ip}(t) - s_{i,R}(t)$.

(ii) $x_{ip}(t) > \sum_{j=1}^{p-1} x_{ij}(t)$

If all the elements of $\underset{\sim}{x}_i(t)$ *excluding* $x_{ip}(t)$ are missing, then $s_{i,NR}(t) = s_{i,NR}(t-1)\, x_{ip}(t)/x_{ip}(t-1)$. If some of the elements of $\underset{\sim}{x}_i(t)$ excluding $x_{ip}(t)$ are missing, the choice of $s_{i,NR}(t)$ is not so obvious. In any event, one must have that $s_{i,R}(t) + s_{i,NR}(t) < x_{ip}(t)$. To this end, four separate possible imputations for $s_{i,NR}(t)$ will be given in order of preference.

(a) $s_{i,NR}(t) = [s_{i,NR}(t-1) + s_{i,R}(t-1)]\, x_{ip}(t)/x_{ip}(t-1) - s_{i,R}(t)$ provided that $s_{i,NR}(t) \geq 0$. Note that the condition $x_{ip} > \Sigma_{j=1}^{p-1} x_{ij}(t)$ is met if $s_{i,NR}(t) \geq 0$.

(b) $s_{i,NR}(t) = s_{i,NR}(t-1)\, [s_{i,R}(t)/s_{i,R}(t-1)]$

(c) $s_{i,NR}(t) = s_{i,NR}(t-1)\, [x_{ip}(t)/x_{ip}(t-1)]$

(d) $s_{i,NR}(t) = x_{ip}(t) - s_{i,R}(t)$.

The preferred imputation will be the first one that does not violate the inequality condition. For all the above cases, the imputed (actual values) will then be

$$I_{ij}^{(2)}(t) = (1-\delta_{ij})\, [s_{i,NR}(t)/s_{i,NR}(t-1)]\, x_{ij}(t-1)$$
$$+ \delta_{ij}\, x_{ij}(t);\ j=1, \ldots, p-1$$

Case 2: Parts of the elementary vector missing and $x_{ip}(t)$ is missing

As in case 1, two subcases are possible:

(i) $x_{ip}(t) = \sum_{j=1}^{p-1} x_{ij}(t)$

If $\Sigma_{j=1}^{p-1} \delta_{ij} = 0$, then $s_{i,NR}(t) = I_{ip}^{(1)}(t)$ where $I_{ip}^{(1)}(t)$ has been obtained using the imputation for total non-response. The imputation $I_{ij}^{(2)}(t)$ is then used. If $\Sigma_{j=1}^{p-1} \delta_{ij} > 0$, $I_{ij}^{(2)}(t)$ will be used provided that $s_{i,NR}(t) = I_{ip}^{(1)}(t) - s_{i,R}(t) \geq 0$. Otherwise, the following imputation must be used

$$I_{ij}^{(3)}(t) = (1-\delta_{ij})\ [s_{i,NR}(t)/s_{i,NR}(t-1)]\ x_{ij}(t-1)$$
$$+ \delta_{ij}\, x_{ij}(t);\ j=1, \ldots, p-1$$

and $I_{ip}^{(1)}(t)$ is replaced by $I_{ip}^{(3)}(t) = \Sigma_{j=1}^{p-1} I_{ip}^{(3)}(t)$

(ii) $x_{ip}(t) > \Sigma_{j=1}^{p-1} x_{ij}(t)$

For this case, the $x_{ip}(t)$ in case 1(ii) is replaced by $I_{ip}^{(1)}(t)$ and the methods given for this case are used, provided that the above inequality condition is satisfied. If the condition cannot be met, $I_{ip}^{(3)}(t)$ must be used and $I_{ip}^{(1)}(t)$ is replaced by $I_{ip}^{(3)}(t) = \Sigma_{j=1}^{p-1} I_{ip}^{(3)}(t)$.

If the assumption, that the distributions of the data elements of vectors $\underset{\sim}{x}_i(t)$ and $\underset{\sim}{x}_i(t-1)$ is similar, does not hold, then each individual element must be imputed using procedures for imputation for total non-response. These imputations must then be adjusted in order to satisfy the inequality requirement $x_{ip} \geq \Sigma_{j=1}^{p-1} x_{ij}$. Hence, for example, for case 1(i), we would have for $\Sigma_{j=1}^{p-1} \delta_{ij} = 0$,

$$I_{ij}^{(4)}(t) = [x_{ip}(t) / \sum_{j=1}^{p-1} I_{ij}^{(1)}(t)]\ I_{ij}^{(1)}(t)$$

and for $\Sigma_{j-1}^{p-1} \delta_{ij} > 0$

$$I_{ij}^{(4)}(t) = (1-\delta_{ij}) \left[\frac{x_{ip}(t) - \Sigma_{j=1}^{p-1} \delta_{ij}\, x_{ij}(t)}{\Sigma_{j=1}^{p-1}\ (1-\delta_{ij})\ I_{ij}^{(1)}(t)} \right] + \delta_{ij}\, x_{ij}(t);\ j = 1, \ldots, p-1.$$

Similarly, cases 1(ii) and 2, could be developed using the imputed values $I_{ij}^{(1)}(t)$.

4. CONCLUSION

For periodic business surveys, it is important to have computer systems which can quickly and accurately monitor the flow of in-coming data in terms of its quality. Conversely, for expected data that are not coming in, the system should impute as well as possible for the non-response given some well specified rules.

The editing will cause the flagging of records in possible error. These errors can be termed as critical and non-critical. All errors should be corrected by either reviewing the questionnaires or checking their authenticity with the respondent. If this is not possible on account of time or budgetary constraints, the most critical errors must be corrected. Given that the errors have been taken care of, the next step of the processing is to impute for the non-respondents. Diagnostic summaries of the actions (edits or imputations) taken by the system, should be printed out in order to inform the survey analyst on the status of his data.

REFERENCES

BERTHELOT, J.-M., and HIDIROGLOU, M.A. (1982). Specifications for imputations in the retail trade survey. Technical report, Statistics Canada.

BERTHELOT, J.-M. (1983). Wholesale-retail redesign, statistical edit proposal. Technical Report, Statistics Canada.

DIXON, W.G. (1953). Processing data for outliers. *Biometrics*, 9, 74-89.

ERNST, L.R. (1980). Comparison of estimators of the mean which adjust for large observations. *Sankhya*, 42, 1-16.

FELLEGI, I.P., and HOLT, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

GENTLEMAN, J.F., and WILK, M.B. (1975). Detecting outliers, II. Supplementing the direct analysis of residuals. *Biometrics*, 31, 387-410.

GRUBBS, F.E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11, 1-21.

GUMBEL, E.J. (1960). Discussion on "Rejection of outliers" by Anscombe, F.J. *Technometrics*, 2, 165-166.

HIDIROGLOU, M.A., and SRINATH, K.P. (1981). Some estimators of population totals form a simple random sample containing large units. *Journal of the American Statistical Association*, 76, 690-695.

KENDALL, M.G., and BUCKLAND, W.R. (1957). *A Dictionary of Statistical Terms*. New York: Hafner.

PRESCOTT, P. (1978). Examination of the behaviour of tests for outliers when more than one outlier is present. *Applied Statistics*, 27, 10-25.

SUGAVANAM, R. (1983). A statistical edit for change. Technical Report, Statistics Canada.

SANDE, I.G. (1981). Estimation in the revised ISPI. Technical Report, Statistics Canada.

TIETGEN, G.L., and MOORE, R.H. (1972). Some Grubbs - type statistics for the detection of several outliers. *Technometrics*, 55, 583-598.

WILKINSON, R.G. (1982). An outlier identification technique designed for the Business Finance Annual Survey. Technical Report, Statistics Canada.

Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation

Juan Diego Rodríguez, Aritz Pérez, and
Jose Antonio Lozano, *Member*, *IEEE*

Abstract—In the machine learning field, the performance of a classifier is usually measured in terms of prediction error. In most real-world problems, the error cannot be exactly calculated and it must be estimated. Therefore, it is important to choose an appropriate estimator of the error. This paper analyzes the statistical properties, bias and variance, of the k-fold cross-validation classification error estimator (k-cv). Our main contribution is a novel theoretical decomposition of the variance of the k-cv considering its sources of variance: sensitivity to changes in the training set and sensitivity to changes in the folds. The paper also compares the bias and variance of the estimator for different values of k. The experimental study has been performed in artificial domains because they allow the exact computation of the implied quantities and we can rigorously specify the conditions of experimentation. The experimentation has been performed for two classifiers (naive Bayes and nearest neighbor), different numbers of folds, sample sizes, and training sets coming from assorted probability distributions. We conclude by including some practical recommendation on the use of k-fold cross validation.

Index Terms—k-fold cross validation, prediction error, error estimation, bias and variance, decomposition of the variance, sources of sensitivity, supervised classification.

1 Introduction

Generally, a classifier is induced from training data using a classifier learning algorithm. Each classifier has an associated prediction error, also called the true error. But usually, the true error is unknown, cannot be calculated, and must be estimated from data. This error is called estimated prediction error. An estimator of the error of a classifier is a random variable $\hat{\epsilon}$ and its quality is usually measured by means of its bias and variance. There are several estimators of the classification error, from the simple Resubstitution [8] and Hold-out [21] to the more complex Bootstrap [12] and Bolstered [4]. One of these techniques, and probably the most popular, is k-fold cross validation (k-cv) [26]. In k-cv, the data set is divided into k folds, a classifier is learned using $k-1$ folds, and an error value is calculated by testing the classifier in the remaining fold. Finally, the k-cv estimation of the error is the average value of the errors committed in each fold. Thus, the k-cv error estimator depends on two factors: the training set and the partition into folds.

This paper presents a statistical analysis of the k-cv error estimator focusing on its bias and variance. We propose a novel theoretical decomposition for the variance of k-cv error estimator. The decomposition divides the variance into an irreducible part, independent of the estimator used, and a reducible part, estimator-dependent. Then, the reducible part is divided taking into account the two sources of variance: sensitivity to changes in the training set and sensitivity to changes in the folds. We also compare the bias and variance of the k-cv estimator for different values of k using the Friedman plus Nemenyi hypothesis tests [7]. The study has been performed on artificial domains because they allow the exact computation of the implied quantities and we can specify rigorously the conditions of experimentation.

The rest of the paper is organized as follows: In Section 2, we briefly explain how to estimate the error using k-cv. Section 3 shows the decomposition of the variance. In Section 4, we explain the experimental process and the working out of the experiment. In Section 5, we present the summary of results emphasizing the bias and variance behavior, especially its decomposition. Finally, our conclusions and future work are presented.

2 Estimating the Error Using k-Fold Cross Validation

2.1 Notation and Definitions

A usual approach to *supervised classification* consists of creating a classifier from training data in order to predict the value of a class attribute $C \in \{1,\dots,r\}$, also known as the label, given the predictive attributes or features, $\boldsymbol{X} = (X_1,\dots,X_d)$, of an unseen unlabeled instance $\boldsymbol{x} = (x_1,\dots,x_d)$. This work is focused on discrete domains $X_i \in \{1,\dots,r_i\}$. We suppose that $(\boldsymbol{X}, C)$ is a random vector with a joint feature-label probability distribution $p(\boldsymbol{x},\ c)$.

A *classifier* ψ is a function that maps $\boldsymbol{X}$ into C:

$$\psi: \quad \underset{\boldsymbol{x}}{\{1,\dots,r_1\}\times\cdots\times\{1,\dots,r_d\}} \;\underset{\mapsto c,}{\rightarrow \{1,\dots,r\}}$$

and is learned from a training set $S_n = \{(\boldsymbol{x}^{(1)}, c^{(1)}),\dots,(\boldsymbol{x}^{(n)}, c^{(n)})\}$ with a classifier induction algorithm $A(\cdot)$. Given the induction algorithm $A(\cdot)$, which is assumed to be a deterministic function of the training set, the classifier obtained from a training set S_n is denoted as $\psi = A(S_n)$. In the remainder of this section, we will introduce some notation for a given induction algorithm $A(\cdot)$, and for the sake of brevity, we will omit it from the notation. In the performed experimentation (see Section 4), the induction algorithm used should be clear from the context.

The *prediction error* of a classifier ψ is the probability of wrong classification of unlabeled instances $\boldsymbol{x}$ and is denoted as $\epsilon(\psi)$:

$$\epsilon(\psi) = p(\psi(\boldsymbol{X}) \neq C) = E_{\boldsymbol{X}}[1 - p(\psi(\boldsymbol{x})|\boldsymbol{x})]. \qquad (1)$$

Given $p(x, c)$, the minimum theoretical prediction error is given by the *Bayes classifier* [6], [23], ψ_B, which is defined as:

$$\psi_B(\boldsymbol{x}) = \underset{c}{argmax}\{p(c|\boldsymbol{x})\} = c_B(\boldsymbol{x}).$$

We define the *Bayes error* as the prediction error of the Bayes classifier:

$$\epsilon(\psi_B) = E_{\boldsymbol{X}}[1 - p(c_B(\boldsymbol{x})|\boldsymbol{x})] = \sum_{\boldsymbol{x}} (1 - p(c_B(\boldsymbol{x})|\boldsymbol{x})) \cdot p(\boldsymbol{x}).$$

Note that this error does not depend on training data or sample size since the Bayes classifier depends only on the feature-label probability distribution of the domain. Any other classifier has a higher than or equal error as the Bayes classifier.

Nevertheless, in most real-world problems, the feature-label probability distribution is unknown. So, both the Bayes classifier and its prediction error are unknown. Moreover, the prediction error of a classifier ψ is also unknown, cannot be exactly computed, and thus, must be estimated. In order to analyze the estimated error, it is necessary to consider the concepts of bias and variance of the estimator used. Let ϵ be the real error of the classifier and $\hat{\epsilon}$ the estimation of the error. The **bias** of an error estimator is defined as the real error value minus the expected estimated error value ($\epsilon - E[\hat{\epsilon}]$). An estimator is said to be **unbiased** if it has zero bias. The **variance** of an error estimator is given by $E[(\hat{\epsilon} - E[\hat{\epsilon}])^2]$.

Intuitively, the bias measures the average precision of the error estimation, while the variance measures the variability of the estimation of the error.

- *The authors are with the Intelligent Systems Group, Computer Science Faculty, University of the Basque Country (UPV-EHU), Paseo Manuel de Lardizabal 1, E-20018 Donostia—San Sebastián, Gipuzkoa, Spain. E-mail: {juandiego.rodriguez, aritz.perez, ja.lozano}@ehu.es.*

Manuscript received 15 Dec. 2008; revised 12 June 2009; accepted 22 July 2009; published online 18 Nov. 2009.
Recommended for acceptance by M. Meila.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-12-0857.
Digital Object Identifier no. 10.1109/TPAMI.2009.187.

0162-8828/10/$26.00 © 2010 IEEE Published by the IEEE Computer Society

2.2 k-Fold Cross-Validation Error Estimator

In k-cv, a data set S_n is uniformly at random partitioned into k folds of similar size $P = \{P_1, \ldots, P_k\}$. For the sake of clarity and without loss of generality, we will suppose that n is multiple of k. Let $T_i = S_n \setminus P_i$ be the complement data set of P_i. Then, the algorithm $A(\cdot)$ induces a classifier from T_i, $\psi_i = A(T_i)$, and estimates its prediction error with P_i. The k-cv prediction error estimator of $\psi = A(S_n)$ is defined as follows [26]:

$$\hat{\epsilon}_k(S_n, P) = \frac{1}{n} \sum_{i=1}^{k} \sum_{(\boldsymbol{x},c) \in P_i} 1(c, \psi_i(\boldsymbol{x})), \tag{2}$$

where $1(i, j) = 1$ iff $i \neq j$ and zero otherwise. So, the k-cv error estimator is the average of the errors committed by the classifiers ψ_i in their corresponding partitions P_i. The estimated error can be considered a random variable which depends on the training set S_n and the partition P.

Generally, an estimator $\hat{\epsilon}$ is a *randomized error estimator* if there are internal random factors that affect its outcome. On the other hand, if the error estimator is a deterministic function, it is a *nonrandomized error estimator* and its variance due to internal factors is zero [6]. For example, k-cv with $k < n$ is a randomized error estimator because it depends on the partition P used, and k-cv with $k = n$ is deterministic because there is no randomness, as there is only one possible partition of the data.

A k-cv error estimator is an unbiased estimator of the prediction error ϵ on data sets of $n - n/k$ size [2], but it is biased for ϵ on data sets of size n because only a subset of the instances with size $n - n/k$ is used for training. This is called *the surrogate problem* [5]. Intuitively, this characteristic will cause k-cv to be a pessimistic estimator. On the other hand, with regard to the variance, it is known that there is no unbiased estimator of the variance $Var[\hat{\epsilon}_k(S_n, P)]$ of k-cv [1].

The *repeated* m *times* k*-cv* (m-k-cv) consists of estimating the error as the average of mk-cv estimations with different random partitions $\boldsymbol{P} = \{P^{(1)}, \ldots, P^{(m)}\}$:

$$\hat{\epsilon}_{k,m}(S_n, \boldsymbol{P}) = \frac{1}{m} \sum_{i=1}^{m} \hat{\epsilon}_k(S_n, P^{(i)}).$$

It is supposed [17] that the repeated version stabilizes the error estimation, and therefore, it reduces the variance of the k-cv estimator, especially for small samples, but, as far as we know, no proof has been given.

As can be deduced from the previous definitions, when a classifier induction algorithm $A(\cdot)$ is fixed, k-cv and m-k-cv estimators have two sources of variance (when $k < n$). One comes from the training sets S_n used for the training test process and the other comes from the partition P (or partitions $\boldsymbol{P}$) of S_n because it affects the internal training test partitions. So, the k-cv and m-k-cv estimators are sensitive to changes in both the training set and the partitions. But what part of the total variance depends on the estimator used and what part is independent? How are the different sources of variance defined and how are they related with the total variance? What is their relative importance for determining the total variance? So, as to answer these interesting questions, the next section provides a novel decomposition of the variance.

3 Decomposition of the Variance of the k-cv Estimator

In order to analyze the behavior of the variance of cross validation, we use the following random variables. All of these variables are defined given a classifier induction algorithm $A(\cdot)$ and a probability $p(\boldsymbol{x}, c)$. The true prediction error random variable ϵ measures the prediction error of a classifier induced with $A(\cdot)$, and follows the distribution $p(\epsilon = e) = \sum_{S_n | \epsilon(S_n) = e} p(S_n)$ (see (1)). The estimated error random variable $\hat{\epsilon}_k$ measures the estimated prediction error of a classifier induced with $A(\cdot)$ by means of the k-cv procedure and follows the distribution $p(\hat{\epsilon}_k = e) = \sum_{S_n, P | \hat{\epsilon}_k(S_n, P) = e} p(S_n, P)$ (see (2)). Note that $p(S_n, P) = p(S_n)p(P)$ due to the independence of S_n and P. The deviation of the error random variable δ_k measures the deviation $\delta_k(S_n, P) = \epsilon(S_n) - \hat{\epsilon}_k(S_n, P)$ and follows the distribution $p(\delta_k = e) = \sum_{S_n, P | \delta_k(S_n, P) = e} p(S_n, P)$.

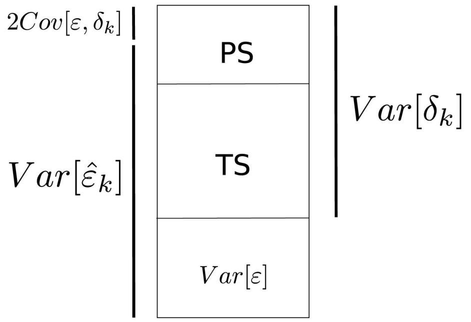


Fig. 1. Decomposition of the variance of k-cv estimator.

The estimated error $\hat{\epsilon}_k$ can be written as $\hat{\epsilon}_k = \epsilon - \delta_k$. Thus, its variance can be decomposed into three terms:

$$Var[\hat{\epsilon}_k] = Var[\epsilon] + Var[\delta_k] - 2Cov[\epsilon, \delta_k]. \tag{3}$$

As $\frac{Cov[\epsilon, \delta_k]}{Var[\epsilon]} \longrightarrow_{n \to \infty} 0$, which means that, for big enough n, $Cov_{S_n, P}[\epsilon, \delta_k]$ is negligible compared with $Var_{S_n}[\epsilon]$, we approximate $Var[\hat{\epsilon}_k]$ using the first two terms in (3) (for instance, in our experiments, the covariance is less than 5 percent of the total variance):

$$Var[\hat{\epsilon}_k] \simeq Var[\epsilon] + Var[\delta_k]. \tag{4}$$

Now we can study the variance of the estimation as the variance of the real error (with respect to S_n) plus the variance of the deviation of the error. The variance of the real error ϵ only depends on the training sets used and it is independent of the estimator. We call it *irreducible variance* because it is common to all the estimators. So, in order to study the properties of the k-cv and m-k-cv estimators, it is desirable to subtract it from the total variance $Var[\hat{\epsilon}_k] - Var[\epsilon] = Var[\delta_k]$. The variance of δ_k is the variance of the precision of the estimation. It is the part of the total variance associated with the estimator used and we call it *reducible variance*. It depends on both the training sets S_n and the partitions P used.

The variance of δ_k can be decomposed into exactly two terms (see the Appendix), depending on the sources of variability, i.e., training and partition sensitivity:

$$Var[\delta_k] = TS + PS, \tag{5}$$

where TS and PS summarize the sensitivities due to changes in the training sets and changes in the partitions, respectively. The definition of both terms is as follows:

$$TS = 1/2(Var_{S_n}[E_P[\delta_k]] + E_P[Var_{S_n}[\delta_k]]), \tag{6}$$

$$PS = 1/2(Var_P[E_{S_n}[\delta_k]] + E_{S_n}[Var_P[\delta_k]]), \tag{7}$$

where $Var_{S_n}[\cdot]$, $E_{S_n}[\cdot]$ and $Var_P[\cdot]$, $E_P[\cdot]$ are the variances and expectations with respect to the distribution of S_n and P, respectively.

A representation of the overall decomposition can be seen in Fig. 1.

4 Experimental Study

In this section, we empirically study the statistical properties of the k-cv estimator, bias, and variance, and analyze the variance using

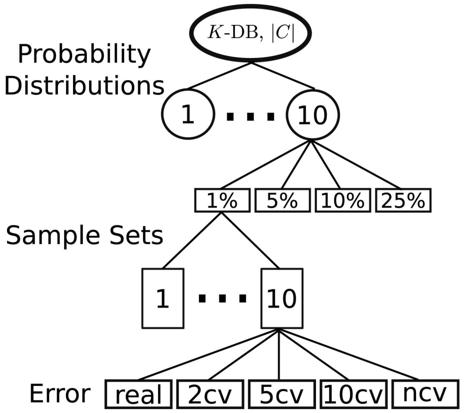


Fig. 2. The creation of the artificial data sets and the estimation of the error.

the decomposition proposed in the previous section. First, we present the artificial domains and the classifiers that we have used, and subsequently, the empirical process and the obtained results.

4.1 Artificial Domains

We use artificial data sets because it allows us to calculate the real prediction error instead of using the empirical one. For this purpose, we sample the data sets from artificial feature-label probability distributions represented as Bayesian networks [24]. As the probability distributions are artificial, we are able to control their complexity and make the experimentation in a wide scenario. To that end, we have used K-dependence Bayesian classifier (K-DB) [25] structures because they allow us to control the number of dependencies among features. A K-DB structure allows each predictive variable X_i to have a maximum of K dependencies with other predictive variables, and in this paper, we have chosen the following K values: $\{0, 1, 2, 3\}$. When the value of K is fixed to 0, it is called the naive Bayes [19], [22] and when K is fixed to 1, it is called the forest-augmented naive Bayes (*FAN*) [20].

4.2 Naive Bayes and K-NN Classifiers

The experimentation includes the study of the k-cv estimator for two different classifiers: naive Bayes (nB) [19], [22] and nearest neighbor (NN) [9]. We have decided to use these classifiers due to their opposite and extreme nature from the point of view of the number of parameters required for each model. A classifier with a high number of parameters can fit the training set very well, with the risk of overfitting, and be very sensitive to changes in it. On the other hand, limiting the number of parameters in order to avoid overfitting reduces the flexibility of the model to capture trends in the data, and can reduce its sensitivity [3], [11]. After introducing both paradigms, we briefly analyze the number of parameters required by them in order to establish their relative sensitivities.

The nB classifier can be considered as a Bayesian network with a special graph topology. It assumes that the predictive variables are conditionally independent given the class, the class being the only parent of each predictor variable. In order to obtain the a posteriori probability distribution of the class given the predictors $p(c|\boldsymbol{x})$, it uses the Bayes rule:

$$p(c|\boldsymbol{x}) = \frac{p(c, \boldsymbol{x})}{p(\boldsymbol{x})} \propto p(c, \boldsymbol{x}).$$

The factorization of the joint probability is very simple because of its independence assumption:

$$p(c, \boldsymbol{x}) = p(c) \prod_{i=1}^{d} p(x_i|c).$$

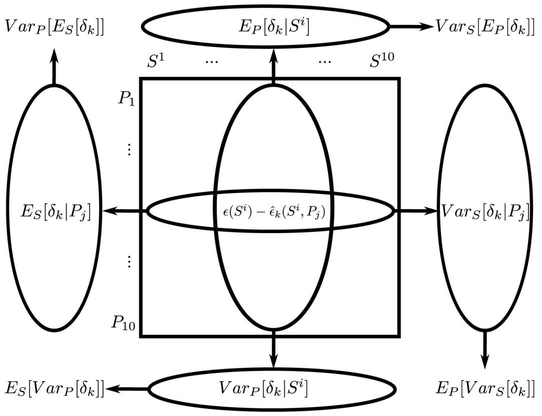


Fig. 3. Computation of training and partition sensitivity.

Generally, nB classifies a new case $\boldsymbol{x}$ using the a posteriori distribution together with the *winner-takes-all* rule:

$$c^* = \underset{c}{argmax}\{p(c|\boldsymbol{x})\} = \underset{c}{argmax}\{p(c, \boldsymbol{x})\}.$$

The nB classifier requires $r - 1 + \sum_{i=1}^{d}(r_i - 1) \cdot r$ parameters, where r is the cardinality of the class variable, r_i is the cardinality of the predictive variable X_i, and d is the number of predictive attributes. The low number of parameters needed by nB is due to the strong conditional independence of each pair of predictive variables given the class variable. It should be noted that the number of parameters needed is independent of the number of instances n in the training set.

The NN classifier is based on a distance measure. In order to classify a new instance, it computes the distances to every case in the training set and then selects the class which belongs to the nearest case. The NN classifier requires $n \cdot (d + 1)$ parameters so that, considering that, in our experiments, $n \gg d$, the number of parameters of NN is higher than the number of parameters of nB. NN is known as a lazy classifier because it does not construct an explicit model of the data from the training set and needs to store all of the available data, if a case condensed or selection technique is not performed.

It is generally accepted that the error estimation of a classifier has higher variance and lower bias as the number of required parameters increases, or equivalently, as the sensitivity to the changes in the training sets increases [3], [11].

4.3 The Empirical Process

We consider domains with 10 predictive attributes and one class attribute. The predictive attributes are binary and the class attribute cardinality ranges from 2 to 5. In order to obtain assorted distributions with different dependencies and complexity degrees, the procedure in Fig. 2 has been carried out. For each K of K-DB and class cardinality, we generate 10 random distributions encoded with the previously described Bayesian networks. Then, for each generated Bayesian classifier, we sample 10 data sets of each sample size. The selected sample sizes are 1, 5, 10, and 25 percent of the total size of the probability space.

This empirical process is summarized in Fig. 2. In total, 6,400 data sets are generated (four different K values, four different class cardinalities, 10 distributions for each class cardinality and K value, four different sample sizes, and 10 sets sampled from each distribution and sample size). Fixed a distribution and a sample size, for each data set $S^1, \ldots, S^{10}$ and each classifier (nB and NN), we estimate 10 times the $\hat{\epsilon}_k(S_n, P)$ for 10 different random data partitions $P_1, \ldots, P_{10}$, and 10 times the $\hat{\epsilon}_{k,m}(S_n, \boldsymbol{P})$ for 10 different sets of random partitions $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_{10}$, where $\boldsymbol{P}_i = (P_i^1, \ldots, P_i^{10})$.

Then, departing from the previous calculated values and exact values, we estimate the variance and expected values of the

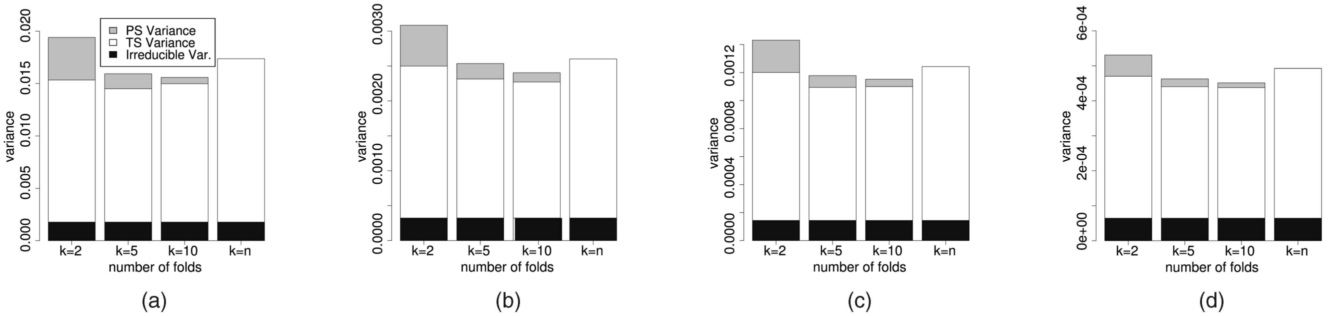


Fig. 4. Variance decomposition on k-cv with naive Bayes classifier. (a) 1 percent. (b) 5 percent. (c) 10 percent. (d) 25 percent.

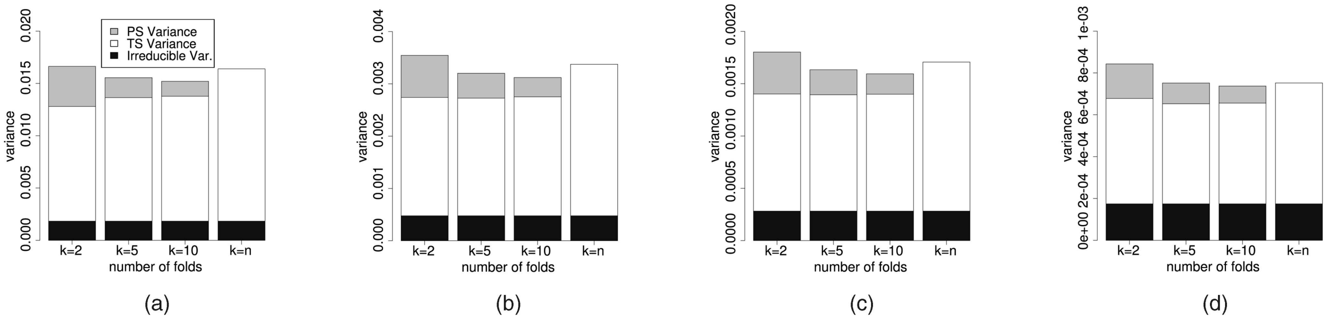


Fig. 5. Variance decomposition on k-cv with nearest neighbor classifier. (a) 1 percent. (b) 5 percent. (c) 10 percent. (d) 25 percent.

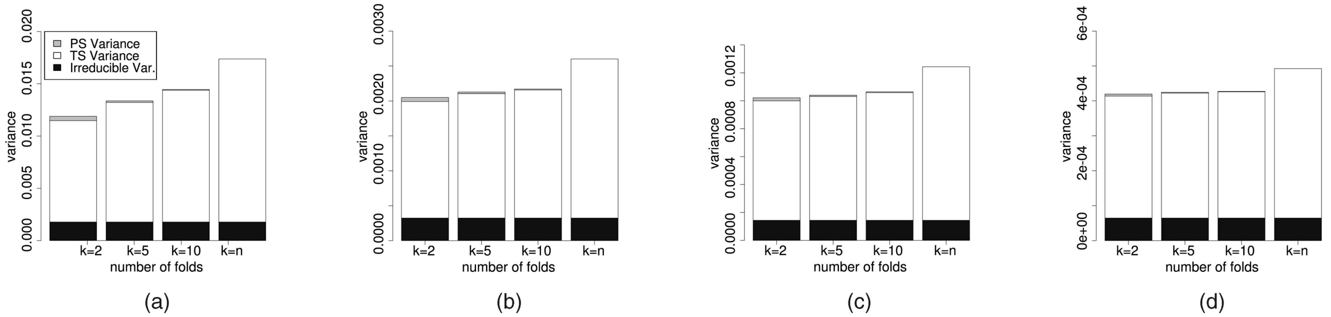


Fig. 6. Variance decomposition on repeated k-cv with naive Bayes classifier. (a) 1 percent. (b) 5 percent. (c) 10 percent. (d) 25 percent.

deviation $\delta_k = \epsilon - \hat{\epsilon}_k$, taking into account the distribution of S_n and P. Finally, we estimate the expected values of the previously estimated variances and the variances of the previously estimated expected values over S_n and P in order to compute TS (6) and PS (7). This process is shown in Fig. 3.

The considered k values for the cross validation are $k = 2, 5, 10, n$. We use the k-cv error estimator provided by the *WEKA* library [28]. The random generated Bayesian networks have been obtained using the *BNGenerator* software [14].

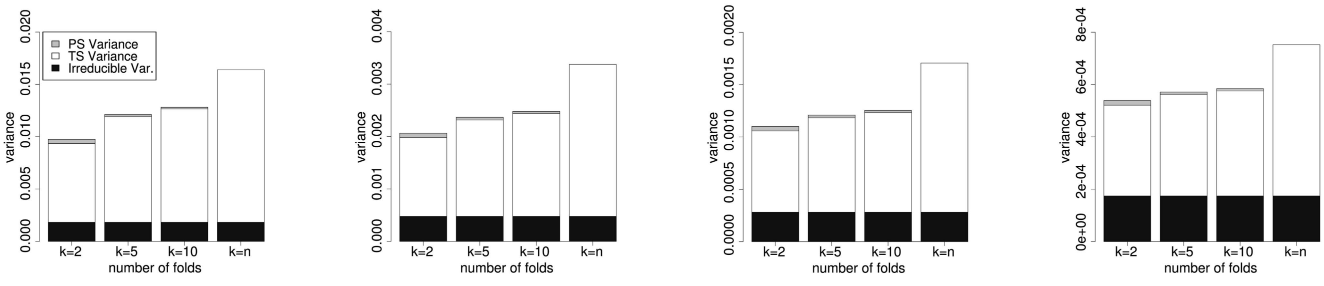
4.4 Experimental Results

This section has been divided into three paragraphs. First, in order to measure the influence of the different sources of variance of the k-cv error estimator, we empirically analyze the decomposition of the variance given in (3). Second, we study the behavior of the bias and the variance of k-cv for different k values and sample sizes using the Friedman plus Nemenyi statistical test [7]. The Friedman test is a nonparametric equivalent of the repeated measures ANOVA. It is used for comparing more than two algorithms over multiple data sets at the same time, based on average ranks. The null hypothesis being tested is that all classifiers obtain the same error. If the null hypothesis is rejected, it can be concluded that there are statistically significant differences between the classifiers, and then, the Nemenyi post hoc test is performed for comparing all classifiers with each other. The results for the Nemenyi post hoc test are shown in critical difference diagrams, and these plots show the mean ranks of each model across all the domains in a numbered line. If there are not statistically significant differences between two classifiers, they are connected in the diagram by a straight line. Finally, we make a brief comparison of the nB and NN classifiers using the Wilcoxon test [7].

4.4.1 *Decomposition of the Variance*

We begin the variance analysis starting out from the decomposition of the variance of the deviation of the error δ_k (5). In Figs. 4, 5, 6, and 7, we present the results of the proposed decomposition (see Fig. 1). Each bar of the figures represents the total variance of the estimator. The lowest, darkest part of the bar, is the irreducible variance: the variance of the true error ϵ. The rest of the bar is the reducible variance, the variance of the deviation of the error δ_k, and is divided into two terms, the sensitivity due to changes in the training set, training sensitivity TS (6), and the sensitivity due to changes in the partitions, partition sensitivity PS (7). Note that PS is zero for $k = n$.

The training sensitivity TS dominates the total variance because it is clearly bigger than the partition sensitivity PS. In nonrepeated k-cv, the training sensitivity TS is 2-4 times bigger with $k = 2$, 4-9 times bigger with $k = 5$, and 5-12 times bigger with $k = 10$. In repeated k-cv, the differences are even greater, the training sensitivity TS is 11-33 times bigger with $k = 2$, 21-80 times



(a) (b) (c) (d)

Fig. 7. Variance decomposition on repeated k-cv with nearest neighbor classifier. (a) 1 percent. (b) 5 percent. (c) 10 percent. (d) 25 percent.

bigger with $k = 5$, and 28-143 times bigger with $k = 10$. In spite of the fact that TS is much bigger than PS, in nonrepeated k-cv, PS is more sensitive than TS for different k values.

In the analysis of the decomposition for different values of k, the partition sensitivity PS decreases with higher values of k. Training sensitivity TS does not have a clear behavior in nonrepeated k-cv, but in repeated k-cv, it increases with higher k values. Finally, it is important to note that the ratio between PS and TS is quite similar for different sizes of the training set, an observation that holds for each k individually.

4.4.2 Comparison of Bias and Variance for Different k Values

In addition to the previous analysis, we have also compared the bias and the total variance of the estimators for the different values of k. In order to do that, we have carried out statistical tests, a paired Friedman test plus the Nemenyi post hoc test when the null hypothesis is rejected [7] based on 320 paired estimated errors (two classifiers, four K values, four class cardinalities, and 10 distributions for each cardinality and K value). The significance of this test is 0.01 (see Figs. 8, 9, 12, and 13).

The first evidence is that, in all of cases, the variance of the estimator decreases with the sample size [6] (see Figs. 10 and 11). Besides, the variance of the estimator is lower on repeated k-cv than in nonrepeated k-cv.

But, there are differences among repeated and nonrepeated k-cv if we focus on the variance for different k-values. In nonrepeated, there are no significative differences between different numbers of folds because the total variance for different k values is very similar (see Fig. 8). Repeated k-cv stabilizes the variance in such a way that significant differences appear (see Fig. 9) and a k ranking from lowest to highest variance arises: $k = 2, 5, 10, n$.

On the other hand, if we focus on the bias, we realize that $k = 2$ is significatively the most biased k value except for nonrepeated k-cv on small samples. The remaining k values show no significative differences among them (see Figs. 12 and 13). The 2-cv has the largest bias for both classifiers (nB and NN) because we use only $n/2$ samples for learning. Anyway, the bias is nearly zero for all sample sizes, specially for sample sizes higher than 5 percent.

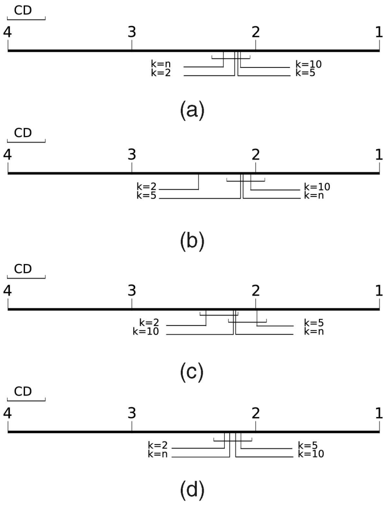
Thus, if the aim is to compare classifiers with similar bias, we should use $k = 2$ because it has the lowest variance. However, if the aim is to measure the prediction error, we should use $k = 5$ or $k = 10$ because they are less biased than $k = 2$ and have less computational cost than $k = n$ (the least biased).

4.4.3 Comparison of nB and NN

Finally, we have also compared the classifiers. The comparison among classifiers (nB and NN) has been performed using the paired Wilcoxon signed-rank test [7], and we have obtained statistically significative results at $\alpha < 0.01$. Table 1 shows the p-values of the statistical tests and the differences between both classifiers. The variance of nB is lower than in NN, especially in nonrepeated k-cv, and NN is less biased than nB due to the differences in the number of parameters [11].

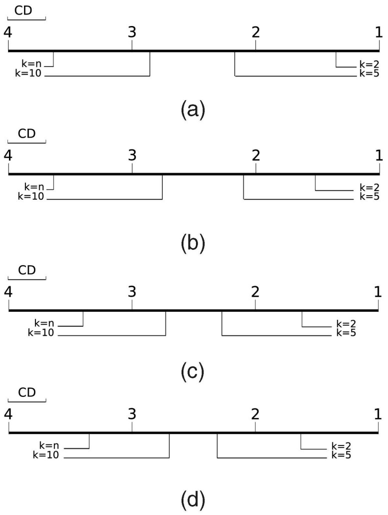
5 Conclusions

This paper proposes a novel decomposition of the variance of the k-fold cross validation for prediction error estimation. The variance is



(a) (b) (c) (d)

Fig. 8. Nemenyi's critical difference diagrams of variance on k-cv. (a) Sample size 1 percent. (b) Sample size 5 percent. (c) Sample size 10 percent. (d) Sample size 25 percent.



(a) (b) (c) (d)

Fig. 9. Nemenyi's critical difference diagrams of variance on repeated k-cv. (a) Sample size 1 percent. (b) Sample size 5 percent. (c) Sample size 10 percent. (d) Sample size 25 percent.

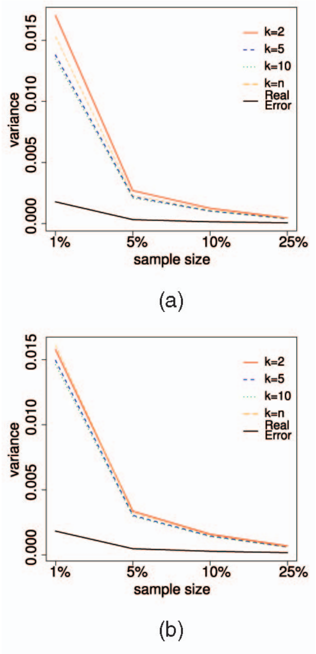


Fig. 10. Variance on k-cv. (a) Naive Bayes. (b) Nearest neighbor.

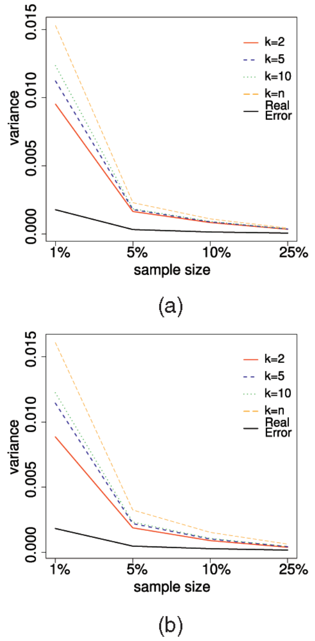


Fig. 11. Variance on repeated k-cv. (a) Naive Bayes. (b) Nearest neighbor.

decomposed into two independent terms (see (4)): the irreducible variance $Var(\epsilon)$ and the reducible variance $Var(\delta_k)$. The irreducible variance is independent of the value of k and the partitions P used, and only depends on the training set. Then, the reducible variance is decomposed into two terms (see (5)) taking into account its sources: the sensitivity due to changes in the training set: TS (see (6)) and due to changes in the partition: PS (see (7)).

Furthermore, the paper empirically studies the statistical properties, bias and variance, of the k-fold cross validation for error estimation and its repeated version. The empirical study is divided into three parts: 1) decomposition of the variance, 2) comparison of bias and variance of the estimator for different k values and training set sizes n, and 3) comparison of bias and variance of the estimator for different induction algorithms, naive Bayes, and nearest neighbor.

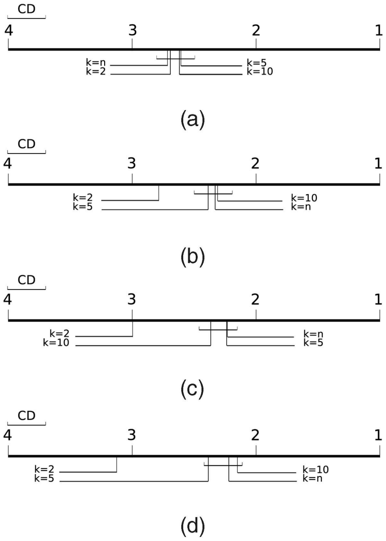


Fig. 12. Nemenyi's critical difference diagrams of bias on k-cv. (a) Sample size 1 percent. (b) Sample size 5 percent. (c) Sample size 10 percent. (d) Sample size 25 percent.

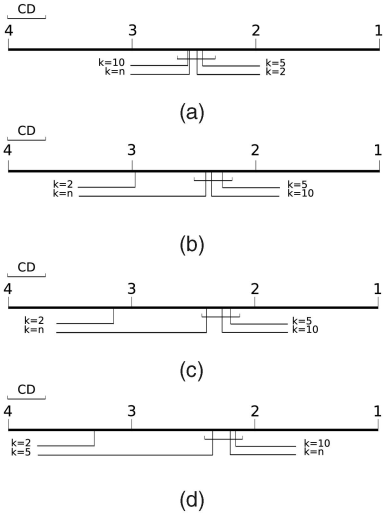


Fig. 13. Nemenyi's critical difference diagrams of bias on repeated k-cv. (a) Sample size 1 percent. (b) Sample size 5 percent. (c) Sample size 10 percent. (d) Sample size 25 percent.

TABLE 1
Wilcoxon Test at $\alpha < 0,01$ between nB and NN Classifiers

k		1%	5%	10%	25%
2	Bias	$\circ 0, 00246$	$\circ 0, 00226$	$\star 0, 00007$	$\star 0, 00006$
	Variance	$\circ 0, 00195$	$\star 0, 00046$	$\star 0, 00093$	$\star 0, 00757$
5	Bias	$\circ 0, 00025$	$\circ 0, 00317$	$\circ 0, 00282$	$\circ 0, 00426$
	Variance	$\star 0, 00128$	$\star 0, 00266$	$\star 0, 00813$	$\star 0, 00814$
10	Bias	$\circ 0, 00061$	$\circ 0, 00326$	$\circ 0, 00321$	$\circ 0, 00487$
	Variance	$\circ 0, 00052$	$\star 0, 00292$	$\star 0, 00049$	$\star 0, 00795$
n	Bias	$\circ 0, 00106$	$\circ 0, 00329$	$\circ 0, 00333$	$\circ 0, 00504$
	Variance	$\circ 0, 00521$	$\star 0, 00120$	$\star 0, 00048$	$\star 0, 00049$

$\star > 0,01\alpha \rightarrow nB < NN$
$\circ > 0,01\alpha \rightarrow NN < nB$

In the first study, we can conclude that training sensitivity TS is much bigger than partition sensitivity PS. PS decreases with higher values of k. TS does not have a clear behavior in nonrepeated k-cv, but in repeated k-cv, TS increases with higher k values. For each k value, the ratio between PS and TS seems to be preserved across different sample sizes. We have observed that the repeated version reduces PS to a small fraction of the total variance. In the second study, we have not found significative differences between the variance of nonrepeated k-cv when the number of folds changes. On the other hand, for 10 times repeated k-cv estimator, a ranking on the variance appears with significant differences between all k values, from the lowest to highest variance: $k = 2, 5, 10, n$. Focusing on the bias, it seems that for k-cv and for repeated k-cv, $k = 2$ is the most biased estimator. In the third study, we realize that NN is less biased than nB but with more variance due to the differences in the number of parameters.

In order to apply these results, we can conclude by recommending $k = 2$ to compare classifiers if their bias is similar because it has the lowest variance. Besides, if the goal is to measure the error, we should use a less biased error estimator. We recommend the use of $k = 5$ or $k = 10$ because they are less biased than $k = 2$ and have less computational cost than $k = n$. Finally, we recommend the use of repeated cross validation when it is computationally feasible.

The theoretical results provided in this paper for k-fold cross-validation estimator could be extended to other error estimators with a part of the variance-dependent of S_n and another part dependent of internal factors, such as Bootstrap.

Appendix

In this section, we demonstrate the exact decomposition of the variance of a random variable Z, which depends on two random independent variables X and Y.

Theorem. *Given two independent random variables X and Y and a third random variable Z, which depends on X and Y, we have that:*

$$\begin{aligned} Var_{X,Y}[Z] = {} & 1/2(E_X[Var_Y[Z]] + Var_Y[E_X[Z]]) \\ & + 1/2(E_Y[Var_X[Z]] + Var_X[E_Y[Z]]). \end{aligned} \tag{8}$$

Proof. By definition of the variance of Z, we have that: $E[(X - EX)^2]$

$$Var_{X,Y}[Z] = E_{X,Y}[Z^2] - E_{X,Y}[Z]^2. \tag{9}$$

We can rewrite this definition by adding and subtracting the term $E_X[E_Y[Z]^2]$ as follows:

$$\begin{aligned} Var_{X,Y}[Z] &= E_{X,Y}[Z^2] - E_X[E_Y[Z]^2] \\ &\quad + E_X[E_Y[Z]^2]) - E_{X,Y}[Z]^2 \\ &= E_X[E_Y[Z^2] - E_Y[Z]^2] \\ &\quad + E_X[E_Y[Z]^2]) - E_X[E_Y[Z]]^2 \\ &= E_X[Var_Y[Z]] + Var_X[E_Y[Z]]. \end{aligned} \tag{10}$$

Following the same procedure with the term $E_Y[E_X[Z]^2]$, we obtain the following equality:

$$Var_{X,Y}[Z] = E_Y[Var_X[Z]] + Var_Y[E_X[Z]]. \tag{11}$$

Using (10) and (11) and regrouping the terms, we prove the theorem

$$\begin{aligned} Var_{X,Y}[Z] &= 1/2(Var_{X,Y}[Z] + Var_{X,Y}[Z]) \\ &= 1/2(E_Y[Var_X[Z]] + Var_Y[E_X[Z]] \\ &\quad + E_X[Var_Y[Z]] + Var_X[E_Y[Z]]) \\ &= 1/2(E_Y[Var_X[Z]] + Var_X[E_Y[Z]]) \\ &\quad + 1/2(E_X[Var_Y[Z]] + Var_Y[E_X[Z]]). \end{aligned}$$

□

We have decomposed the variance of Z into two additive terms which represent the sources of variance due to variables X and Y, respectively (see the two terms of (8)). We call $1/2(E_Y[Var_X[Z]] + Var_X[E_Y[Z]])$ the sensitivity of Z with respect to X, and $1/2(E_X[Var_Y[Z]] + Var_Y[E_X[Z]])$ the sensitivity of Z with respect to Y. This property of the variance allows us to decompose the variance of the estimated prediction error random variable $\hat{\epsilon}_k$ into the sensitivity to changes in the permutation and the sensitivity to changes in the training set.

Acknowledgments

This work has been partially supported by the Saiotek and Research Groups 2007-2012 (IT-242-07) programs (Basque Government), TIN2008-06815-C02-01 and Consolider Ingenio 2010-CSD2007-00018 projects (Spanish Ministry of Science and Innovation), and COMBIOMED network in computational biomedicine (Carlos III Health Institute).

References

[1] Y. Bengio and Y. Grandvalet, “No Unbiased Estimator of the Variance of K-Fold Cross-Validation,” *J. Machine Learning Research,* vol. 5, pp. 1089-1105, 2004.

[2] Y. Bengio and Y. Grandvalet, “Bias in Estimating the Variance of K-Fold Cross-Validation,” *Statistical Modeling and Analysis for Complex Data Problems,* vol. 1, pp. 75-95, 2005.

[3] C.M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[4] U.M. Braga-Neto, R. Hashimoto, E.R. Dougherty, D.V. Nguyen, and R.J. Carroll, “Is Cross-Validation Better than Resubstitution for Ranking Genes?” *Bioinformatics,* vol. 20, no. 2, pp. 253-258, 2004.

[5] U.M. Braga-Neto and E.R. Dougherty, “Is Cross-Validation Valid for Msmall-Sample Microarray Classification?” *Bioinformatics,* vol. 20, no. 3, pp. 374-380, 2004.

[6] U.M. Braga-Neto, “Small-Sample Error Estimation: Mythology versus Mathematics,” *Proc. SPIE,* pp. 304-314, 2005.

[7] J. Demsar, “Statistical Comparisons of Classifiers over Multiple Data Sets,” *J. Machine Learning Research,* vol. 7, pp. 1-30, 2006.

[8] L. Devroye and T. Wagner, “Distribution-Free Performance Bounds with the Resubstitution Error Estimate,” *IEEE Trans. Information Theory,* vol. 25, no. 2, pp. 208-210, Mar. 1979.

[9] L. Devroye, *Non-Parametric Density Estimation.* Wiley, 1985.

[10] P. Domingos, “A Unified Bias-Variance Decomposition and Its Applications,” *Proc. 17th Int'l Conf. Machine Learning,* pp. 231-238, 2000.

[11] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification,* second ed. Wiley Interscience, 2000.

[12] B. Efron and R.J. Tibshirani, “An Introduction to the Bootstrap,” *Monographs on Statistics and Applied Probability,* vol. 57. Chapman and Hall, 1993.

[13] J.H. Friedman, “On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality,” *Data Mining and Knowledge Discovery,* vol. 1, pp. 55-77, 1997.

[14] J.S. Ide and F.G. Cozman, “Generating Random Bayesian Networks with Constraints on Induces Width,” *Proc. European Conf. Artificial Intelligence,* 2004.

[15] G.M. James, “Variance and Bias for General Loss Functions,” *Machine Learning,* vol. 51, pp. 115-135, 2003.

[16] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” *Proc. Int'l Joint Conf. Artificial Intelligence,* pp. 1137-1145, 1995.

[17] R. Kohavi, “Wrappers for Performance Enhancement and Oblivious Decision Graphs,” PhD thesis, Computer Science Dept., Stanford Univ., 1995.

[18] R. Kohavi and D.H. Wolpert, “Bias Plus Variance Decomposition for Zero-One Loss Functions,” *Proc. Int'l Conf. Machine Learning,* pp. 275-283, 1996.

[19] P. Langley, W. Iba, and K. Thompson, “An Analysis of Bayesian Classifiers,” *Proc. 10th Nat'l Conf. Artificial Intelligence,* pp. 223-228, 1992.

[20] P. Lucas, “Restricted Bayesian Network Structure Learning,” *Advances in Bayesian Networks (Studies in Fuzziness and Soft Computing),* pp. 217-232, Springer, 2004.

[21] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley and Sons, Inc., 1992.

[22] M. Minsky, “Steps Toward Artificial Intelligence,” *Trans. IRE,* vol. 49, pp. 8-30, 1961.

[23] T. Mitchell, *Machine Learning.* McGraw-Hill, 1997.

[24] J. Pearl, *Probabilistic Reasoning in Intelligence Systems.* Morgan-Kaufman, 1988.

[25] M. Sahami, “Learning Limited Dependence Bayesian Classifiers,” *Proc. Second Int'l Conf. Knowledge Discovery and Data Mining,* pp. 335-338, 1996.

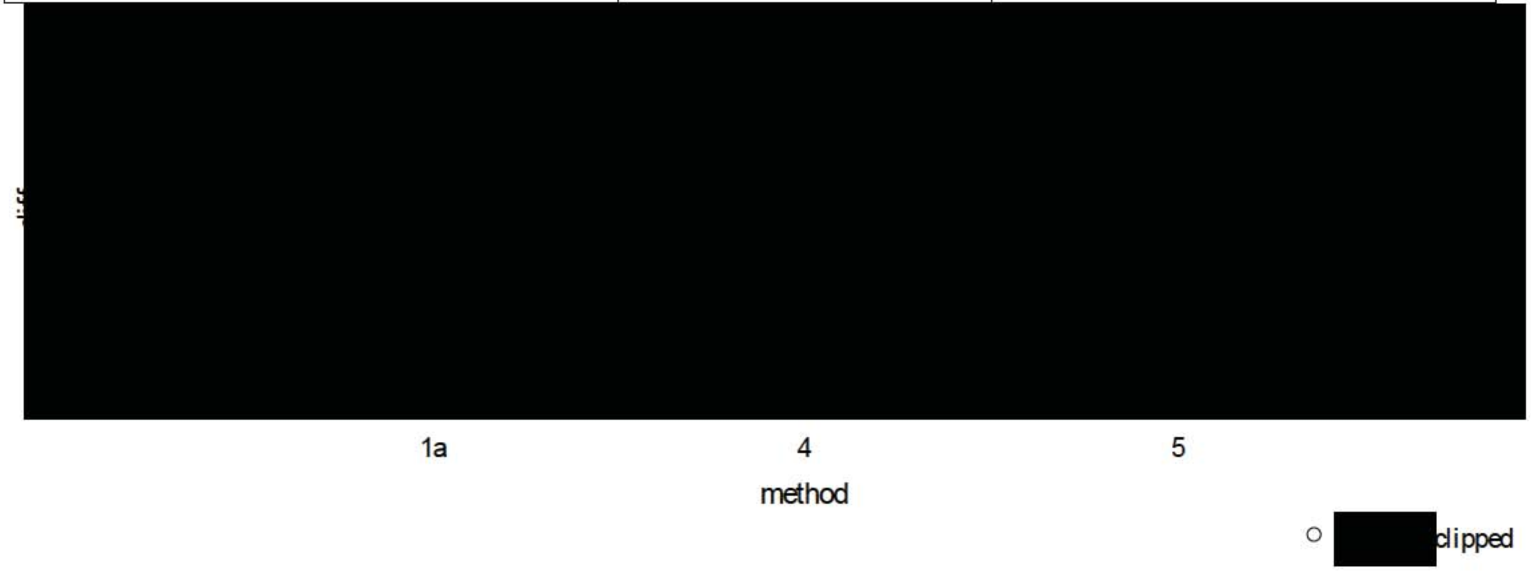
[26] M. Stone, “Cross-Validatory Choice and Assessment of Statistical Predictions,” *J. Royal Statistical Soc. Series B,* vol. 36, pp. 111-147, 1974.

[27] S.M. Weiss and C.A. Kulikowski, *Computer Systems That Learn.* Morgan-Kaufmann, 1991.

[28] I.H. Witten and E. Frank, *Data mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, 2000.

GQTYPCUR=101

Statistics by GQtype			
N	50	900	0
Min	[redacted]	[redacted]	.
Nlow	<15	0	0
Q1	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	.
Mean	-18	74.[redacted]	.
Q3	[redacted]	[redacted]	.
Nhigh	<15	40	0
Max	[redacted]	[redacted]	.
Nout	<15	40	0
Range	850	3900	.
Std Dev	110.[redacted]	248[redacted]	.

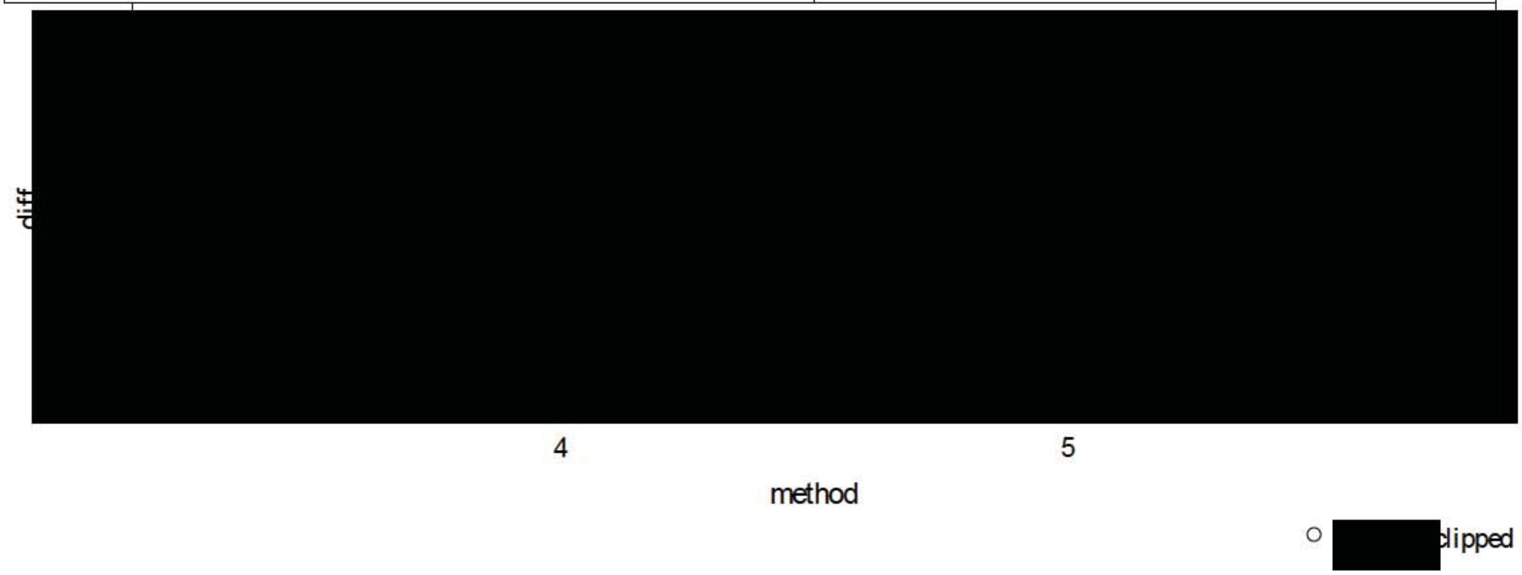


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=102

Statistics by GQtype		
N	250	0
Min	[redacted]	.
Nlow	<15	0
Q1	[redacted]	.
Q2	[redacted]	.
Mean	645[redacted]	.
Q3	[redacted]	.
Nhigh	<15	0
Max	[redacted]	.
Nout	20	0
Range	2800	.
Std Dev	537[redacted]	.

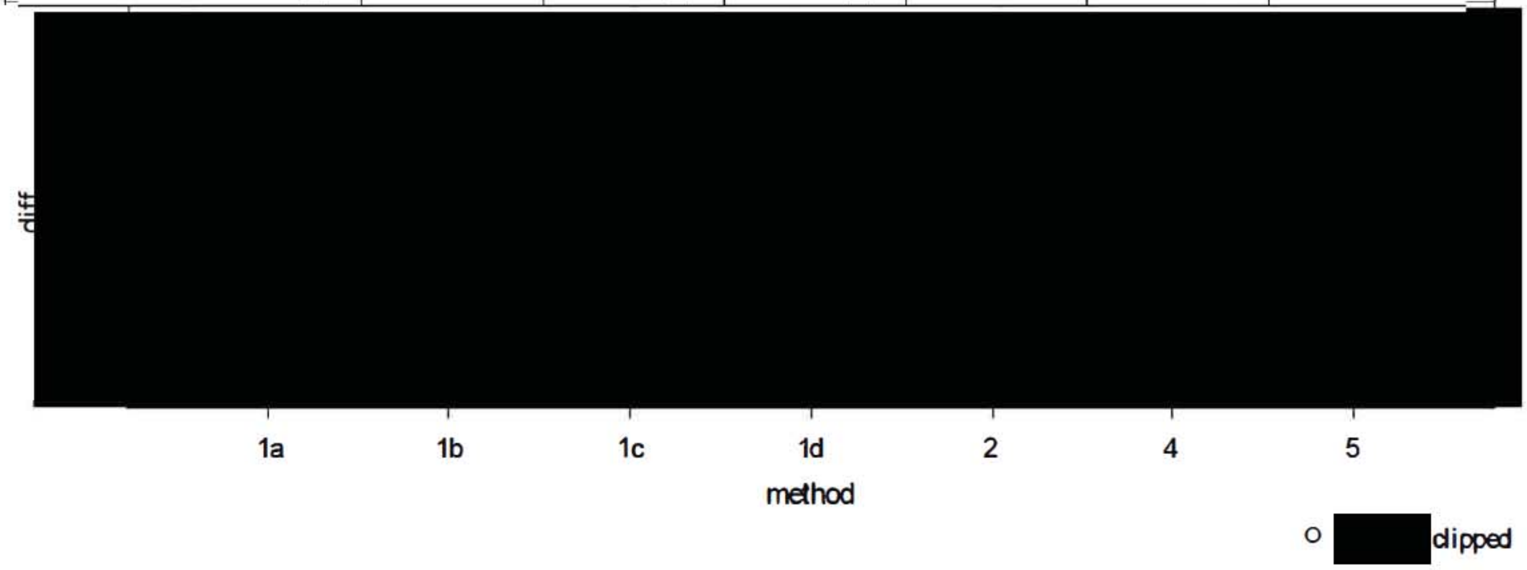

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=103

Statistics by GQtype

N	4700	550	550	550	550	8000	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	250	30	30	30	30	400	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-7.[redacted]	-0[redacted]	-7.[redacted]	-0.[redacted]	-83.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	250	30	30	30	30	400	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	450	50	50	50	50	800	0
Range	3300	1500	2700	1500	1900	4700	.
Std Dev	174.[redacted]	171.[redacted]	170.[redacted]	172.[redacted]	138.[redacted]	199.[redacted]	.

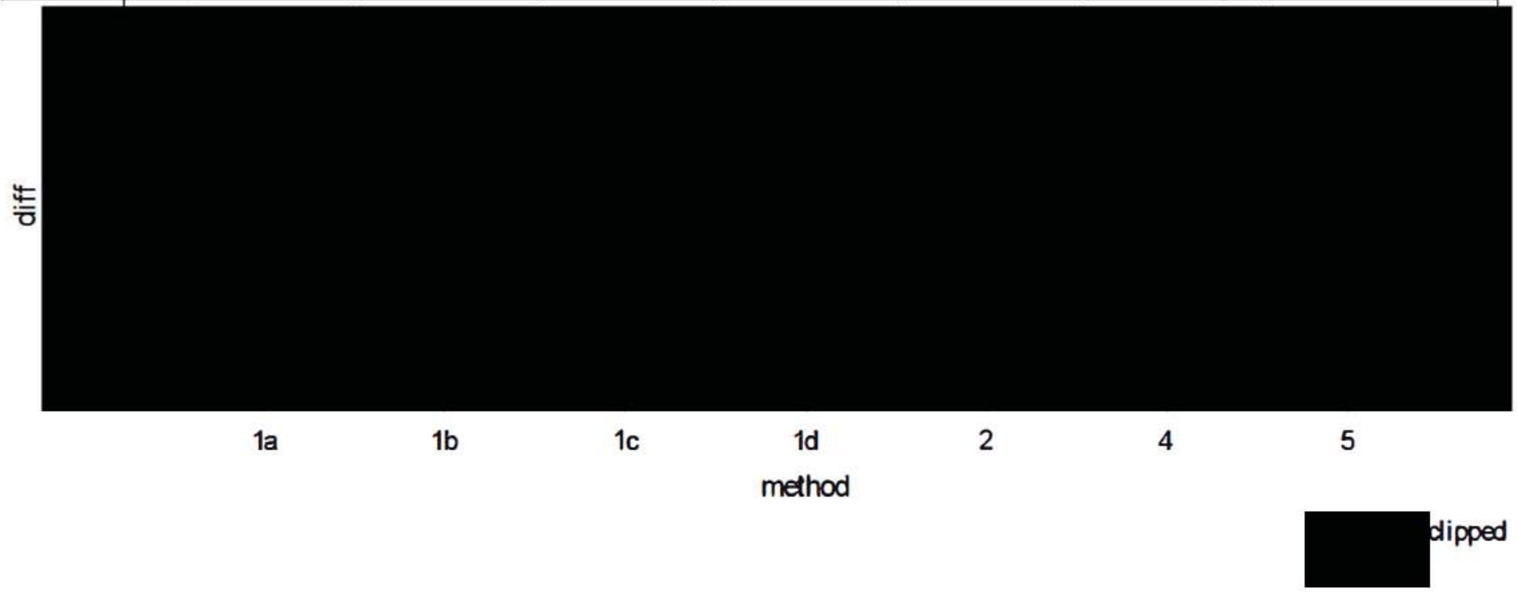

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=104

Statistics by GQtype

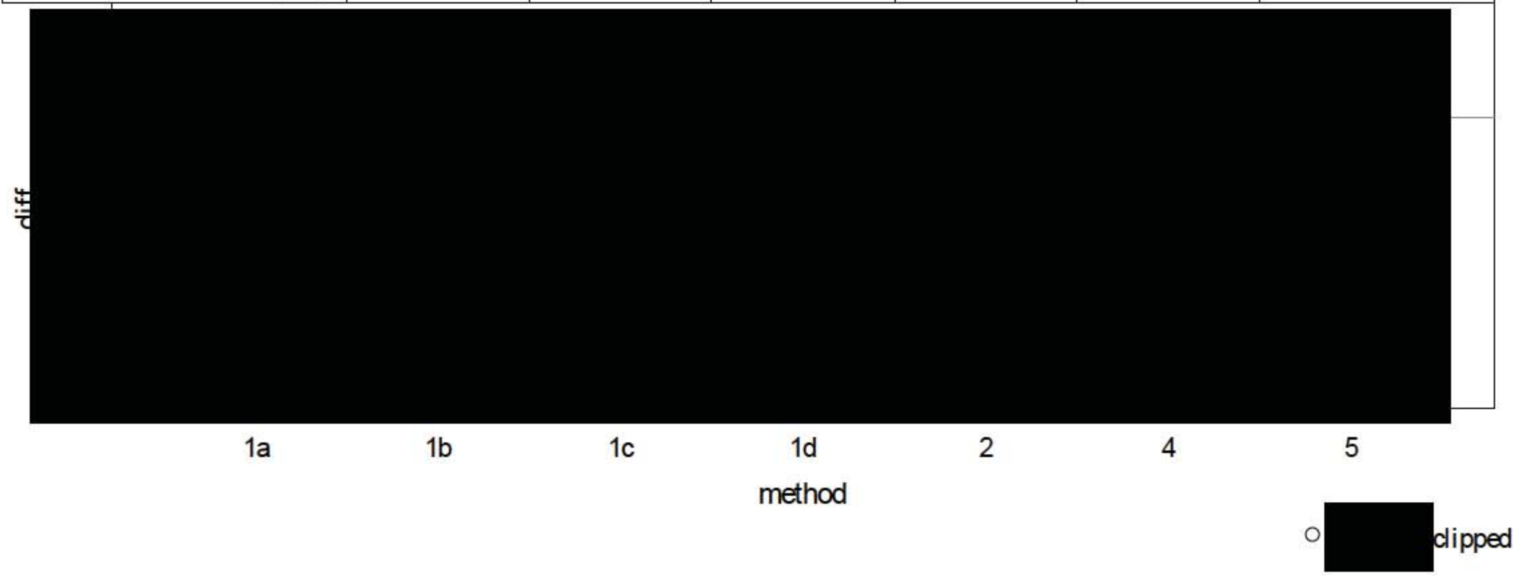
N	2700	2000	2000	2000	2000	2800	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	150	100	100	100	100	150	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	1.[redacted]	-1.[redacted]	-1.[redacted]	-3.[redacted]	-0.[redacted]	-0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	150	100	100	100	100	150	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	250	200	200	200	200	250	0
Range	3000	3600	3100	3600	6000	4100	.
Std Dev	118.[redacted]	136.[redacted]	114.[redacted]	137.[redacted]	145.[redacted]	271.[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=105

Statistics by GQtype							
N	700	350	350	350	350	850	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	30	20	20	20	20	40	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-2.[redacted]	2.[redacted]	-3.[redacted]	-0.[redacted]	0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	30	20	20	20	20	40	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	70	30	30	30	30	80	0
Range	2900	750	600	750	450	3000	.
Std Dev	108.[redacted]	61.[redacted]	63.[redacted]	65.[redacted]	46.[redacted]	164.[redacted]	.

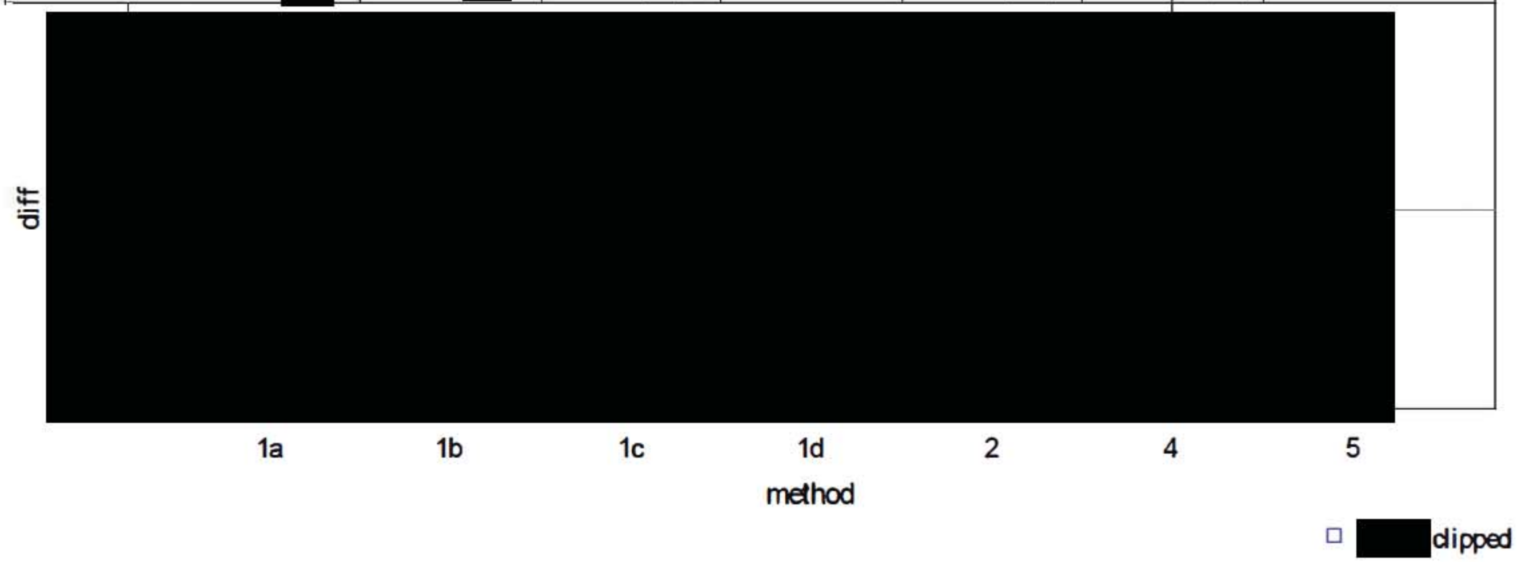

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=106

Statistics by GQtype

N	20	<15	<15	<15	<15	20	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	0	0	0	0	0	0	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-22.[redacted]	-1.[redacted]	24	16[redacted]	-1.[redacted]	-18.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	0	0	0	0	0	0	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	0	0	0	0	0	0	0
Range	250	150	150	100	200	450	.
Std Dev	63.[redacted]	57.[redacted]	55.[redacted]	45.[redacted]	73.[redacted]	136.[redacted]	.

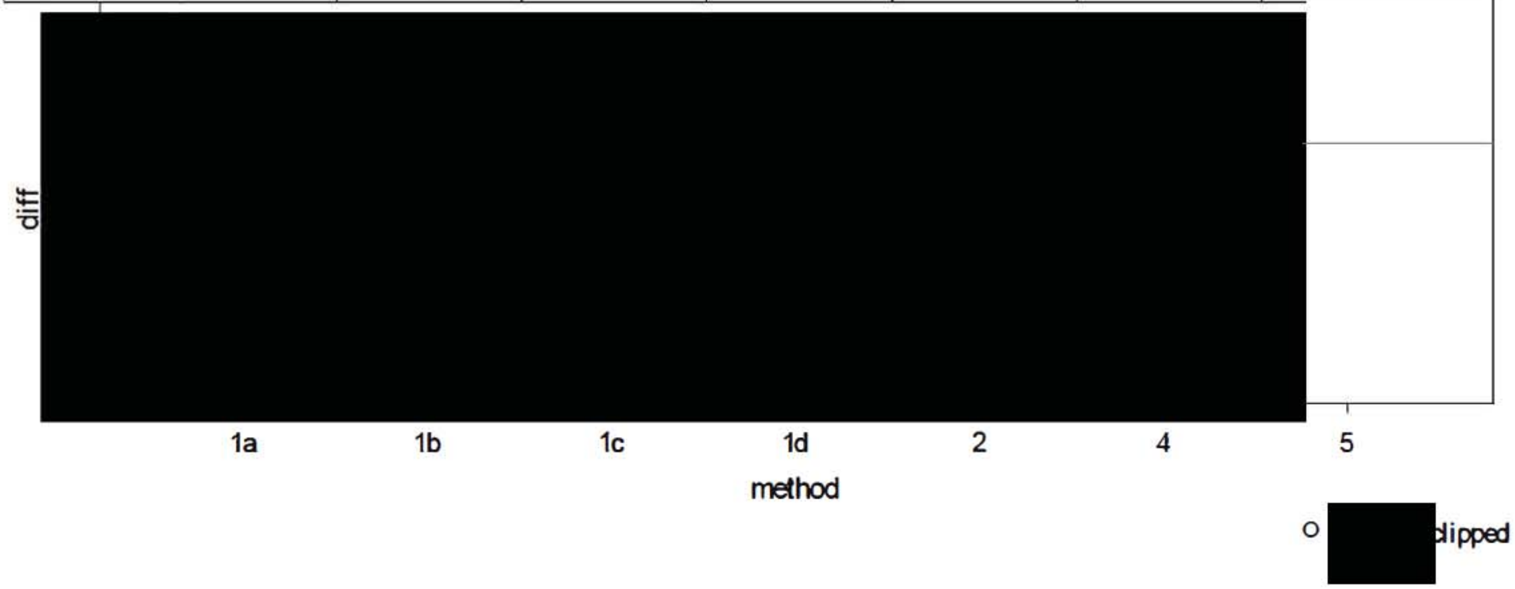

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=201

Statistics by GQtype

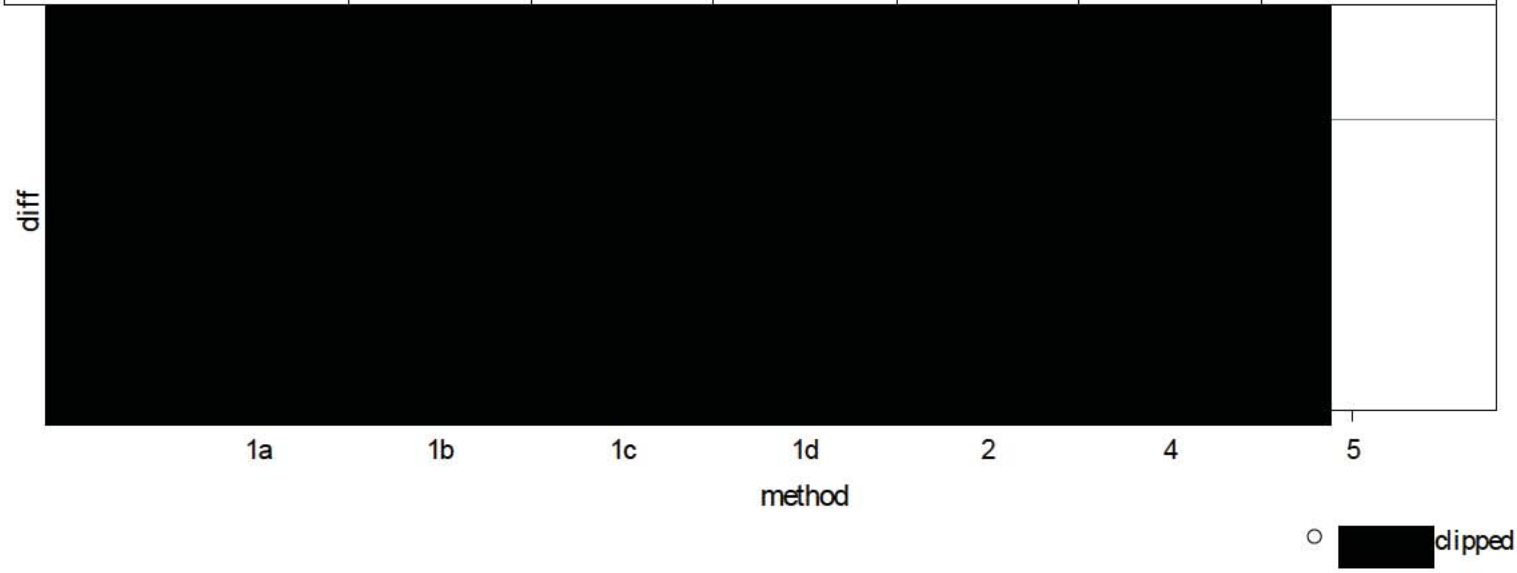
N	3000	1400	1400	1400	1400	3100	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	150	60	70	60	70	100	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-0.[redacted]	-0[redacted]	-0[redacted]	-0.[redacted]	0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	150	70	70	60	70	150	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	300	150	150	100	150	250	0
Range	250	150	150	150	150	200	.
Std Dev	8.[redacted]	9.[redacted]	8.[redacted]	9.[redacted]	8.[redacted]	12.[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=202

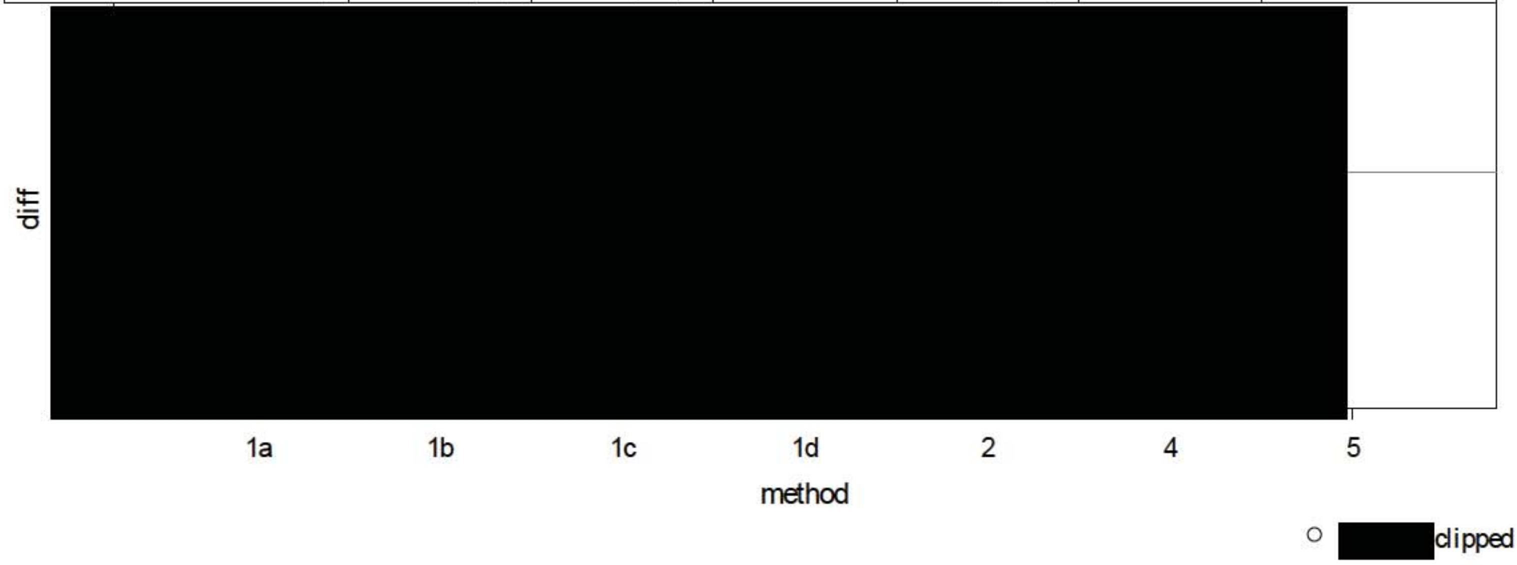
Statistics by GQtype							
N	1600	850	850	850	850	1700	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	80	40	40	40	40	80	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0[redacted]	-2.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	70	40	40	40	40	80	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	150	80	80	80	80	150	0
Range	150	150	200	150	150	300	.
Std Dev	11.[redacted]	13.[redacted]	13.[redacted]	13.[redacted]	11.[redacted]	22.[redacted]	.



GQTYPCUR=203

Statistics by GQtype

N	1000	550	550	550	550	1100	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow							0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-3.[redacted]	-2.[redacted]	-3.[redacted]	-0.[redacted]	0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	50	30	30	30	30	50	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	100	60	60	50	60	100	0
Range	700	650	650	600	250	350	.
Std Dev	26.[redacted]	32.[redacted]	28.[redacted]	30.[redacted]	19.[redacted]	29.[redacted]	.

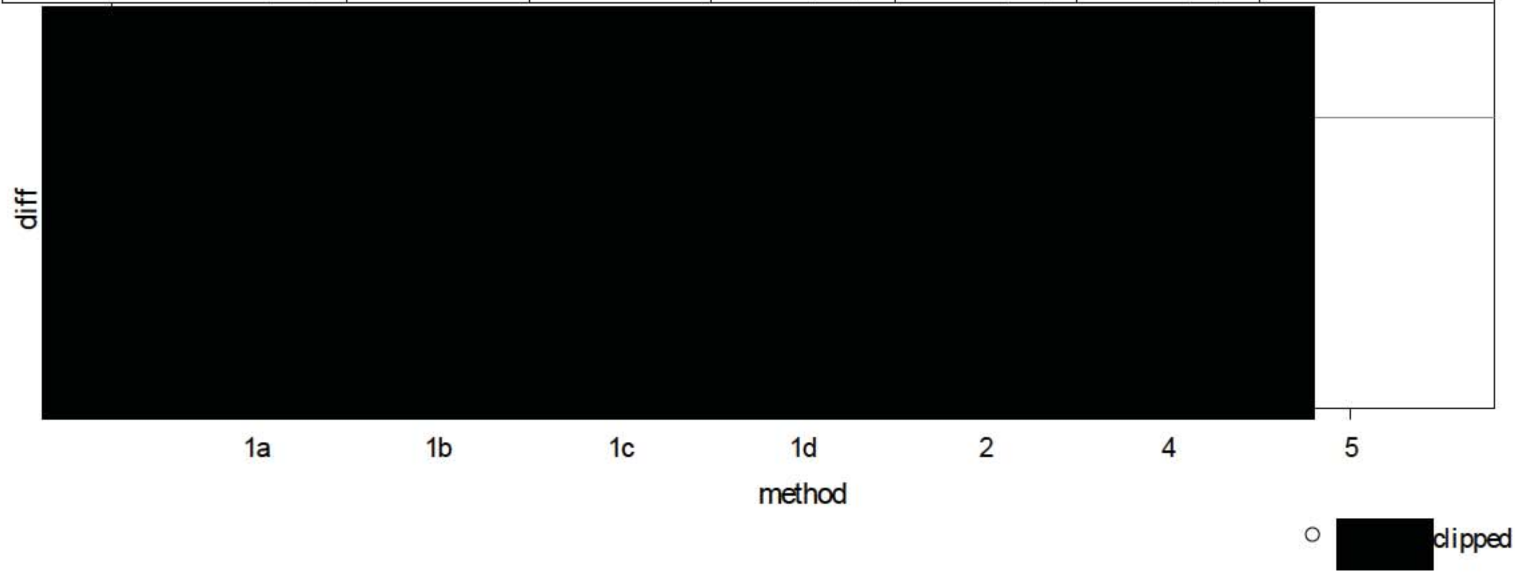

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=301

Statistics by GQtype

N	24000	14500	14500	14500	14500	24500	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	1200	700	700	700	700	1200	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-0.[redacted]	-0.[redacted]	0.[redacted]	-0.[redacted]	-10.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	1200	700	700	700	700	1200	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	2400	1400	1400	1400	1400	2400	0
Range	1400	1500	1100	1500	1100	1400	.
Std Dev	26.[redacted]	27.[redacted]	28.[redacted]	27.[redacted]	25.[redacted]	52.[redacted]	.

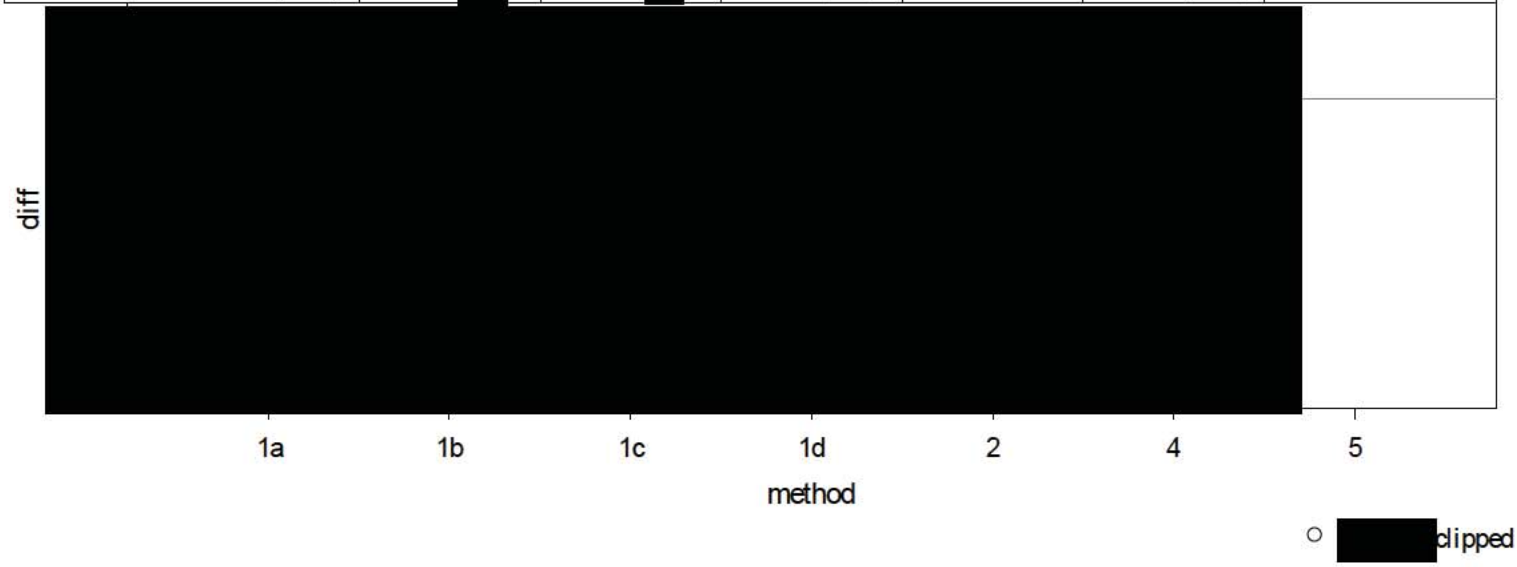

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=401

Statistics by GQtype

N	650	250	250	250	250	700	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	30	<15	<15	<15	<15	30	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-5.[redacted]	-3.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-25.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	30	<15	<15	<15	<15	30	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	60	20	20	20	20	70	0
Range	1400	1400	800	1400	600	1100	.
Std Dev	65.[redacted]	82[redacted]	62.[redacted]	78.[redacted]	46.[redacted]	89.[redacted]	.

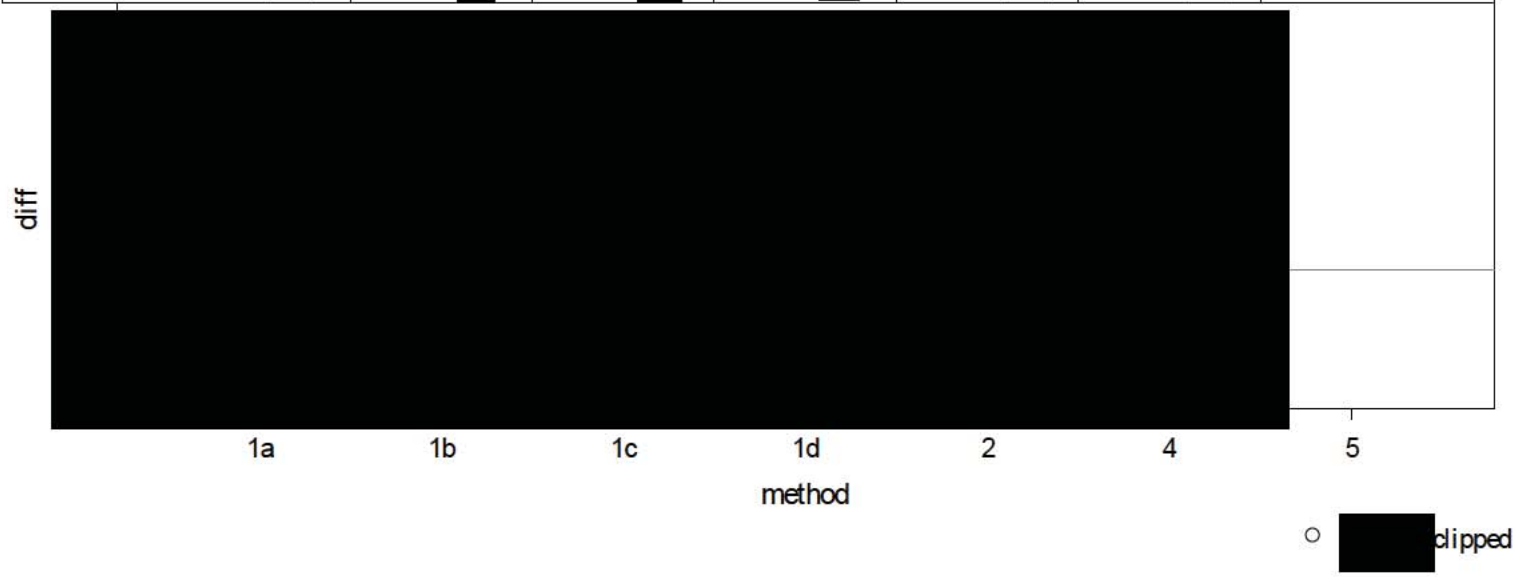

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=402

Statistics by GQtype

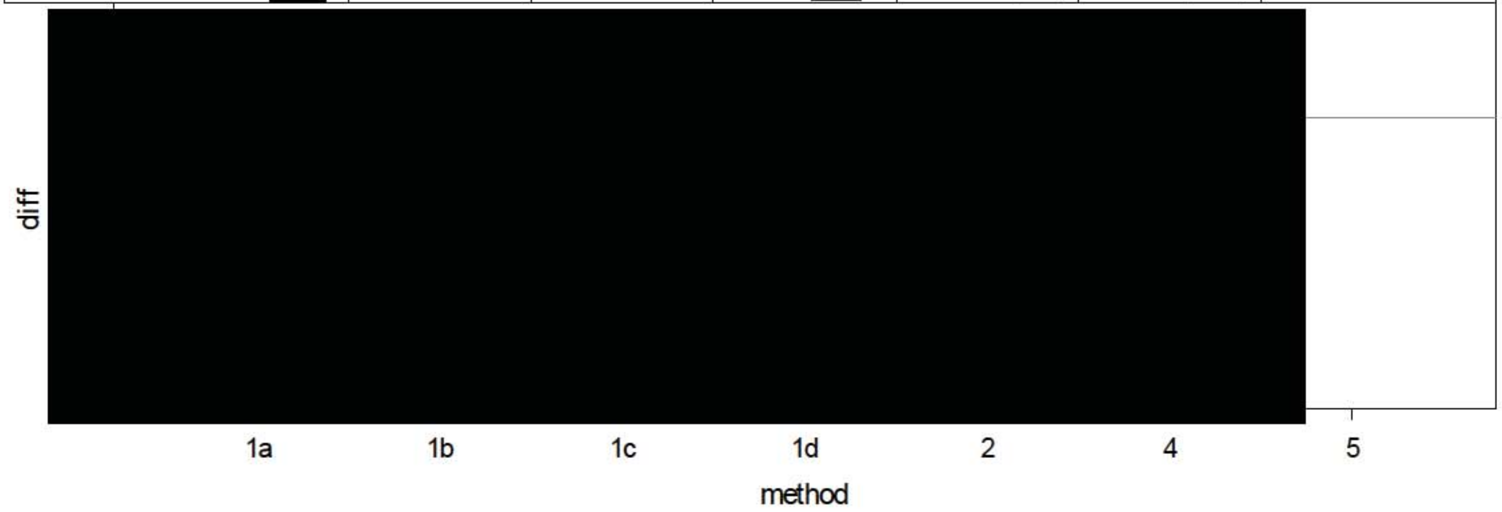
N	200	70	70	70	70	200	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-6.[redacted]	-2.[redacted]	-6[redacted]	-4.[redacted]	-0.[redacted]	-6.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	20	<15	<15	<15	<15	20	0
Range	350	200	200	400	300	600	.
Std Dev	38.[redacted]	26.[redacted]	33.[redacted]	42.[redacted]	30.[redacted]	59.[redacted]	.



DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=403

Statistics by GQtype							
N	450	200	200	200	200	450	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	20	<15	<15	<15	<15	20	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0.[redacted]	-0.[redacted]	-0.[redacted]	0.[redacted]	-0.[redacted]	1[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	20	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	40	20	20	20	20	40	0
Range	150	90	100	90	150	400	.
Std Dev	12.[redacted]	9.[redacted]	11.[redacted]	9.[redacted]	10.[redacted]	39.[redacted]	.


[redacted]

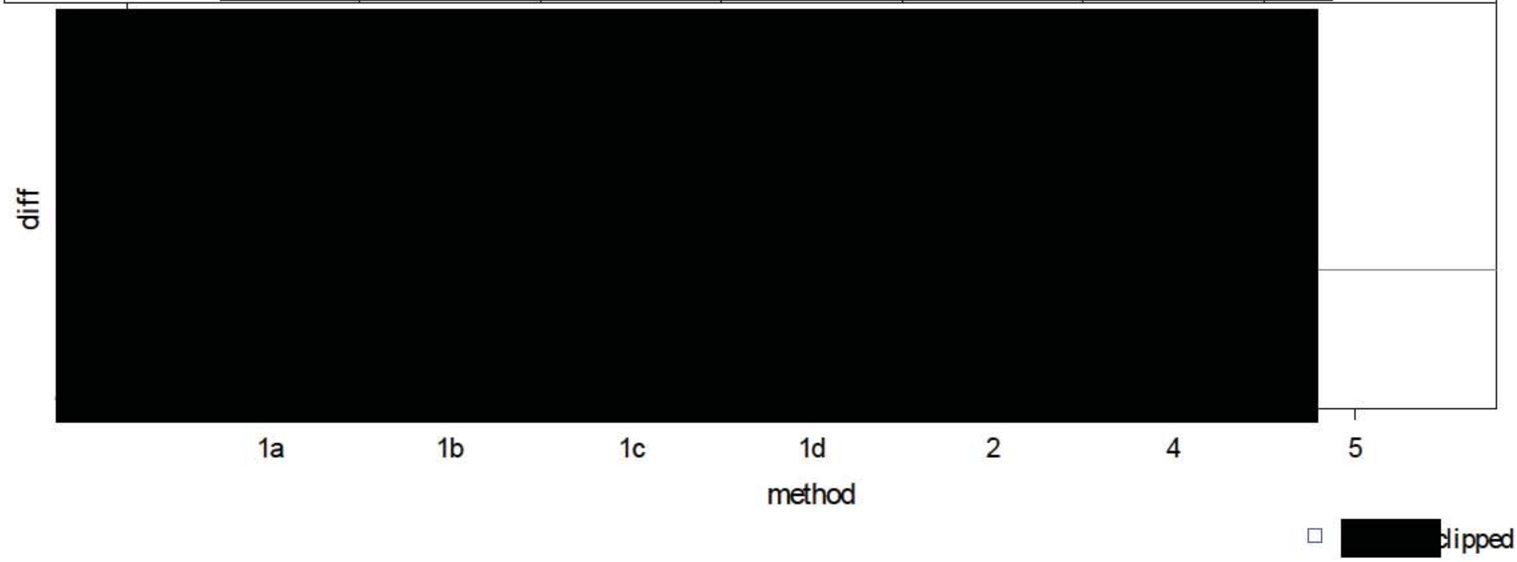
○ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=404

Statistics by GQtype

N	<15	<15	<15	<15	<15	<15	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.

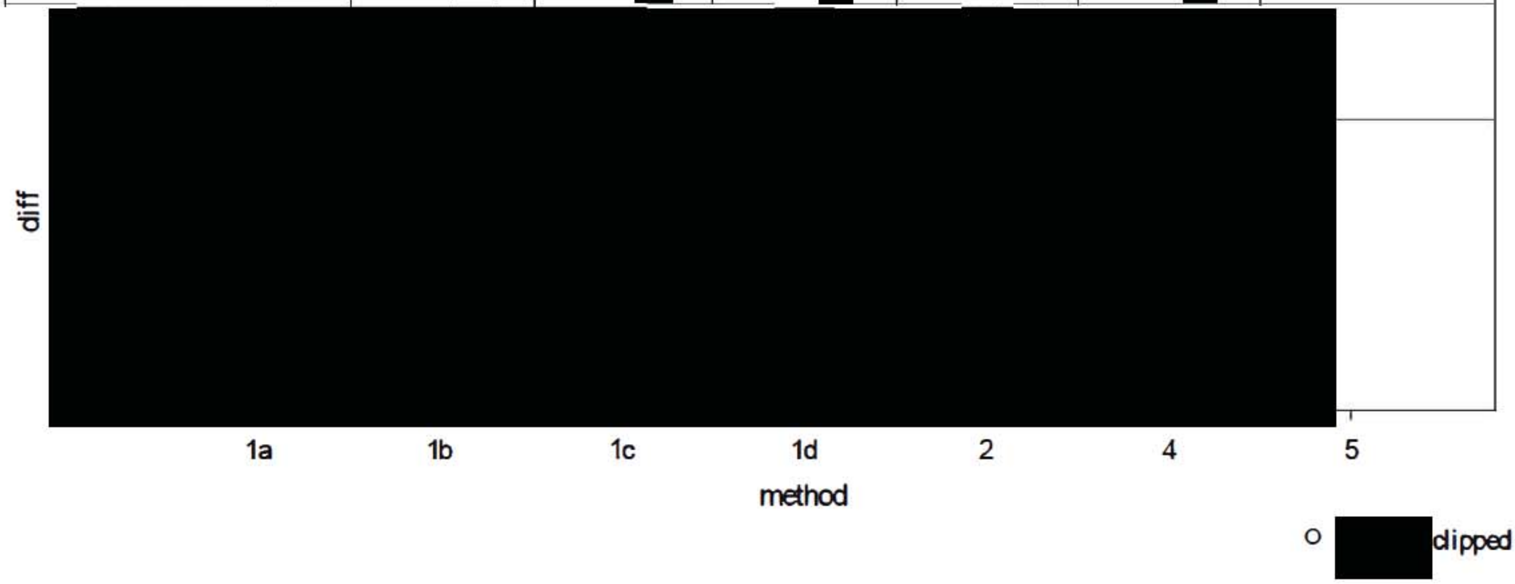

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=405

Statistics by GQtype

N	450	250	250	250	250	500	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	20	<15	<15	<15	<15	20	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1.[redacted]	-3.[redacted]	-0.[redacted]	-3.[redacted]	-0.[redacted]	-6.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	20	<15	<15	<15	<15	20	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	40	20	20	20	20	50	0
Range	350	300	300	300	200	550	.
Std Dev	21.[redacted]	30.[redacted]	25.[redacted]	30.[redacted]	21.[redacted]	39.[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=501

Statistics by GQtype

N	27000	13000	13000	13000	13000	28000	28000
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nlow	1300	650	650	650	650	1300	1400
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	0[redacted]	-1.[redacted]	-0.[redacted]	-1.[redacted]	-0.[redacted]	-10.[redacted]	0.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nhigh	1300	650	650	650	650	1400	1400
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nout	2700	1300	1300	1300	1300	2800	2800
Range	2600	1500	1400	1500	1400	8700	6000
Std Dev	45[redacted]	42.[redacted]	38.[redacted]	42[redacted]	34[redacted]	146.[redacted]	73[redacted]

diff

[redacted]

1a 1b 1c 1d 2 4 5

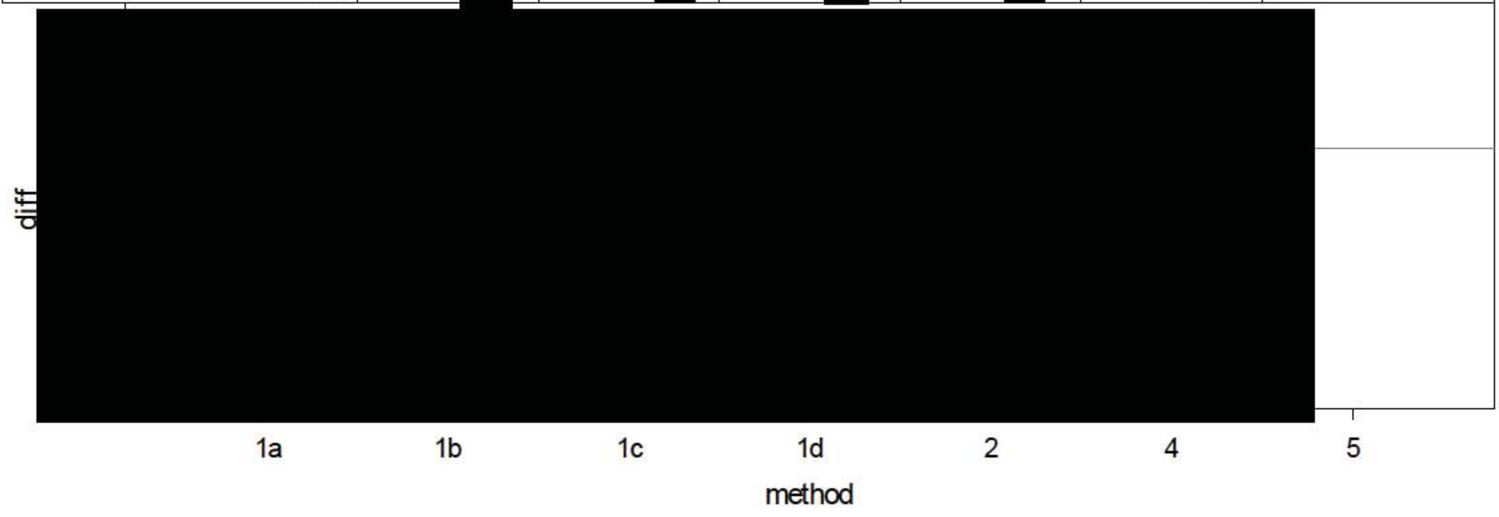
method

○ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=601

Statistics by GQtype							
N	2300	800	800	800	800	2700	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	100	40	40	40	40	100	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-4.[redacted]	-4[redacted]	-4.[redacted]	-1.[redacted]	-0.[redacted]	-33.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	100	40	40	40	40	150	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	250	80	80	80	80	250	0
Range	1900	1400	1300	1400	1200	1700	.
Std Dev	101.[redacted]	95.[redacted]	101.[redacted]	93.[redacted]	82.[redacted]	126.[redacted]	.


[redacted]

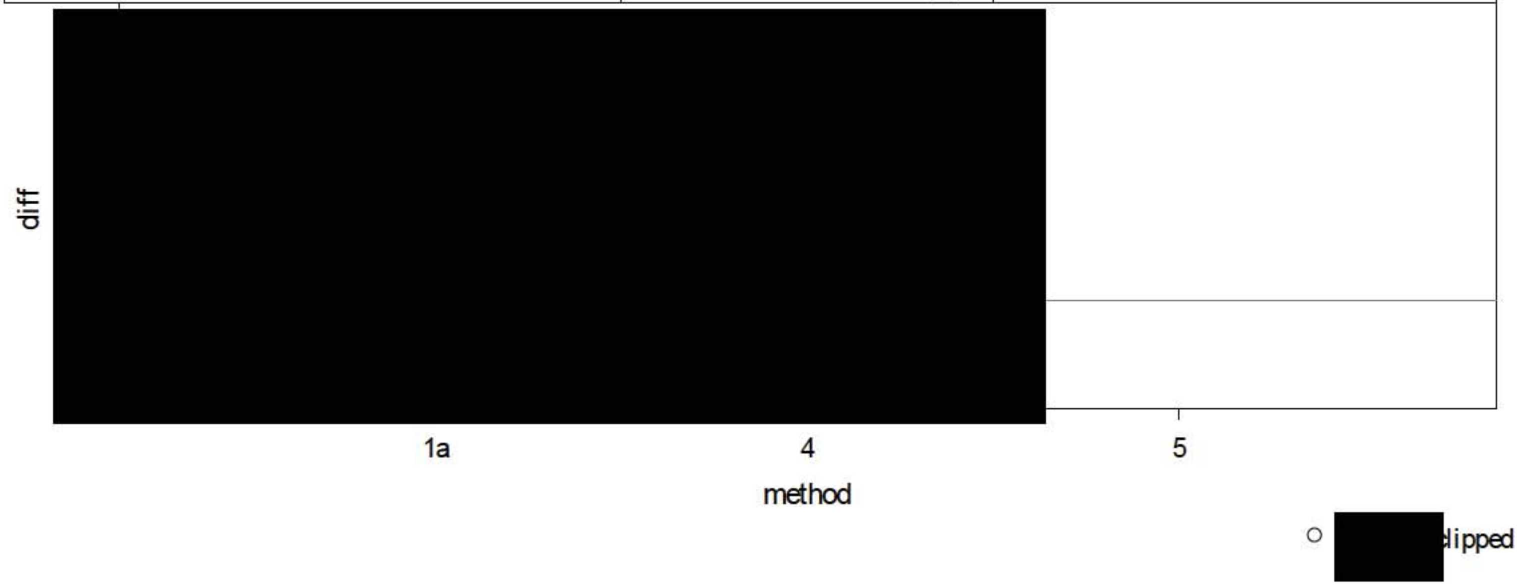
○ [redacted]clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=602

Statistics by GQtype			
N	250	250	0
Min	[redacted]	[redacted]	.
Nlow	<15	<15	0
Q1	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	.
Mean	9	100.[redacted]	.
Q3	[redacted]	[redacted]	.
Nhigh	<15	<15	0
Max	[redacted]	[redacted]	.
Nout	20	20	0
Range	1600	1700	.
Std Dev	135[redacted]	262.[redacted]	.

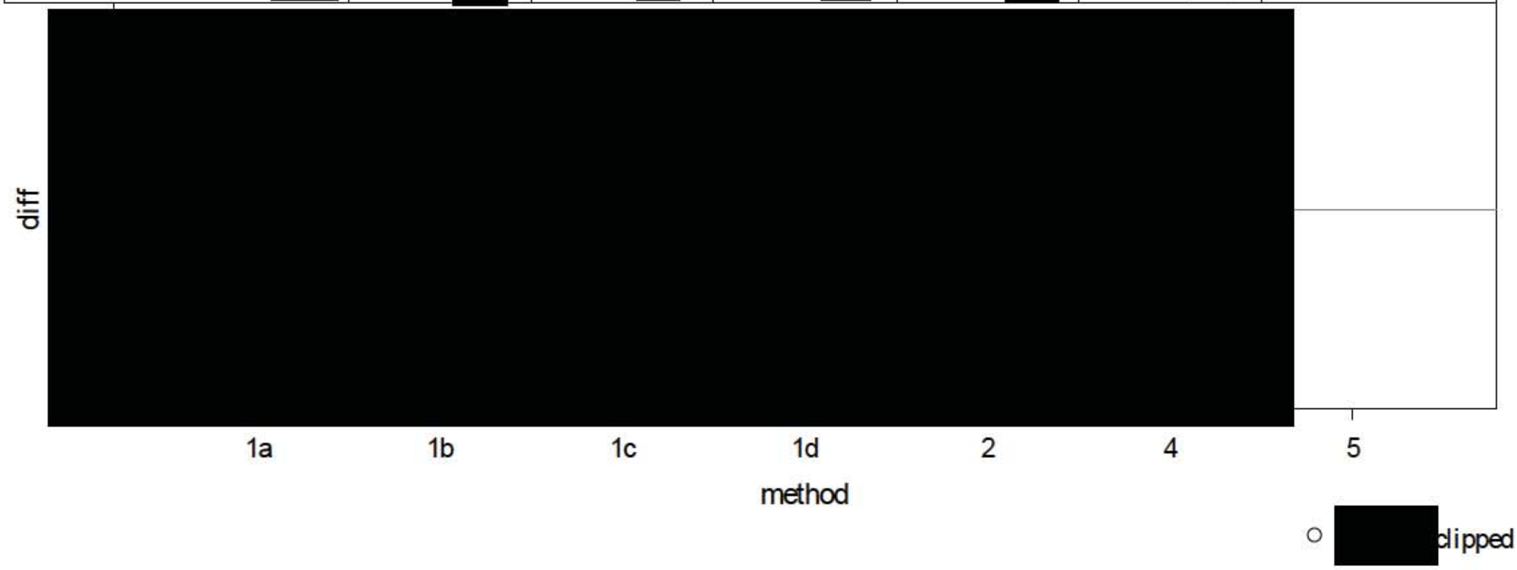
[redacted]



DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=701

Statistics by GQtype							
N	4800	1600	1600	1600	1600	5100	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	250	80	80	80	80	250	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0.[redacted]	-1.[redacted]	-0.[redacted]	-2.[redacted]	-0.[redacted]	5.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	250	80	80	80	80	250	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	450	150	150	150	150	500	0
Range	1300	800	800	800	550	1800	.
Std Dev	48.[redacted]	35.[redacted]	43.[redacted]	36.[redacted]	32.[redacted]	63.[redacted]	.

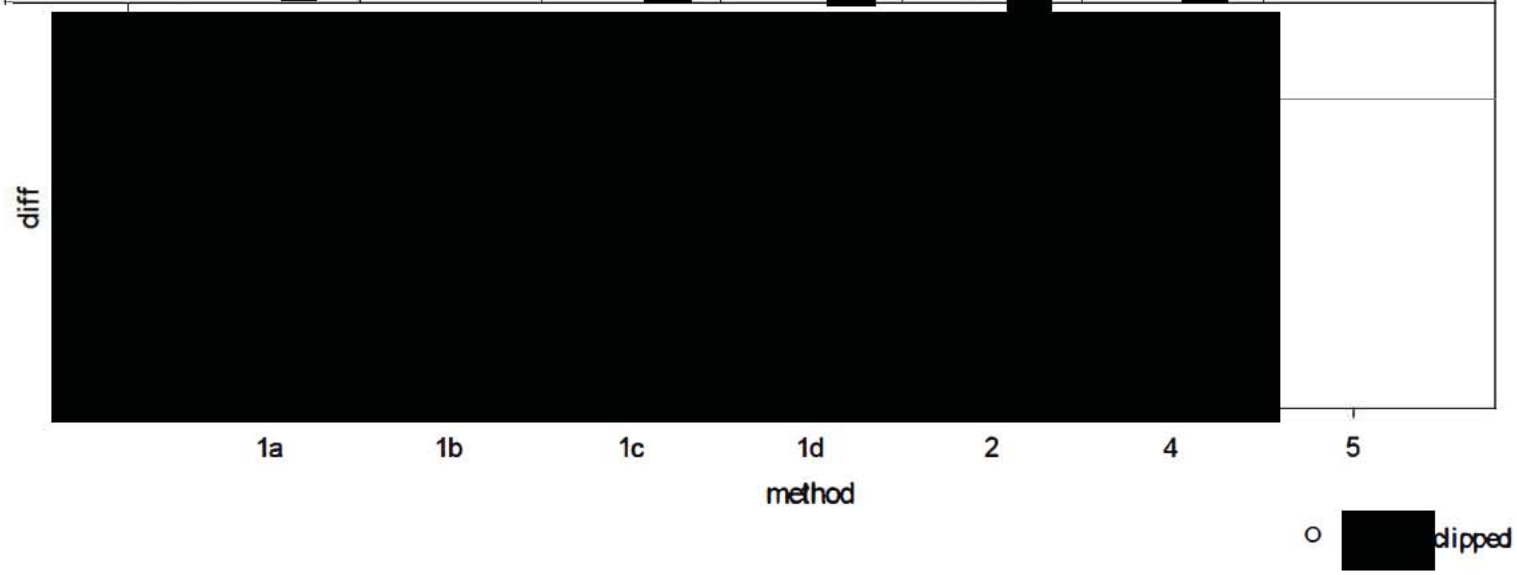


DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=702

Statistics by GQtype

N	1700	700	700	700	700	1800	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	80	40	40	40	40	80	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	4.[redacted]	-7.[redacted]	-2.[redacted]	-12.[redacted]	-0.[redacted]	0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	80	40	30	40	40	90	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	150	70	70	70	70	150	0
Range	1300	650	1700	650	850	1600	.
Std Dev	69.[redacted]	59.[redacted]	82.[redacted]	62.[redacted]	62.[redacted]	89[redacted]	.

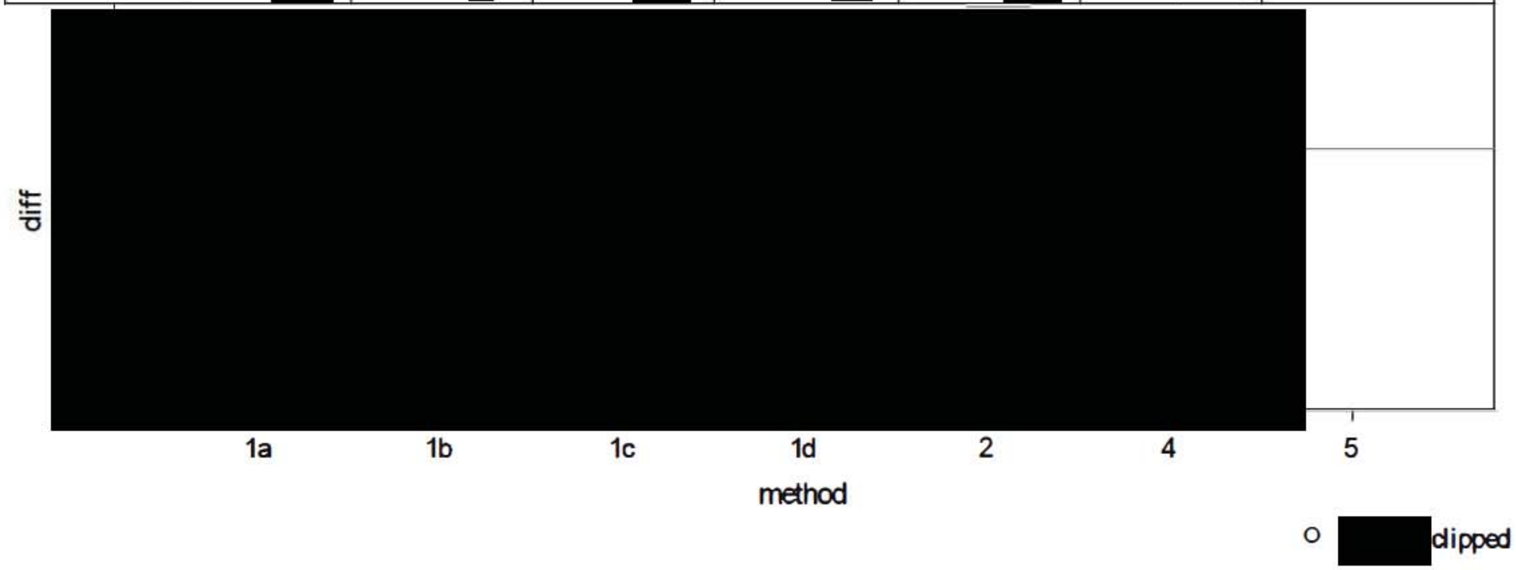

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=704

Statistics by GQtype

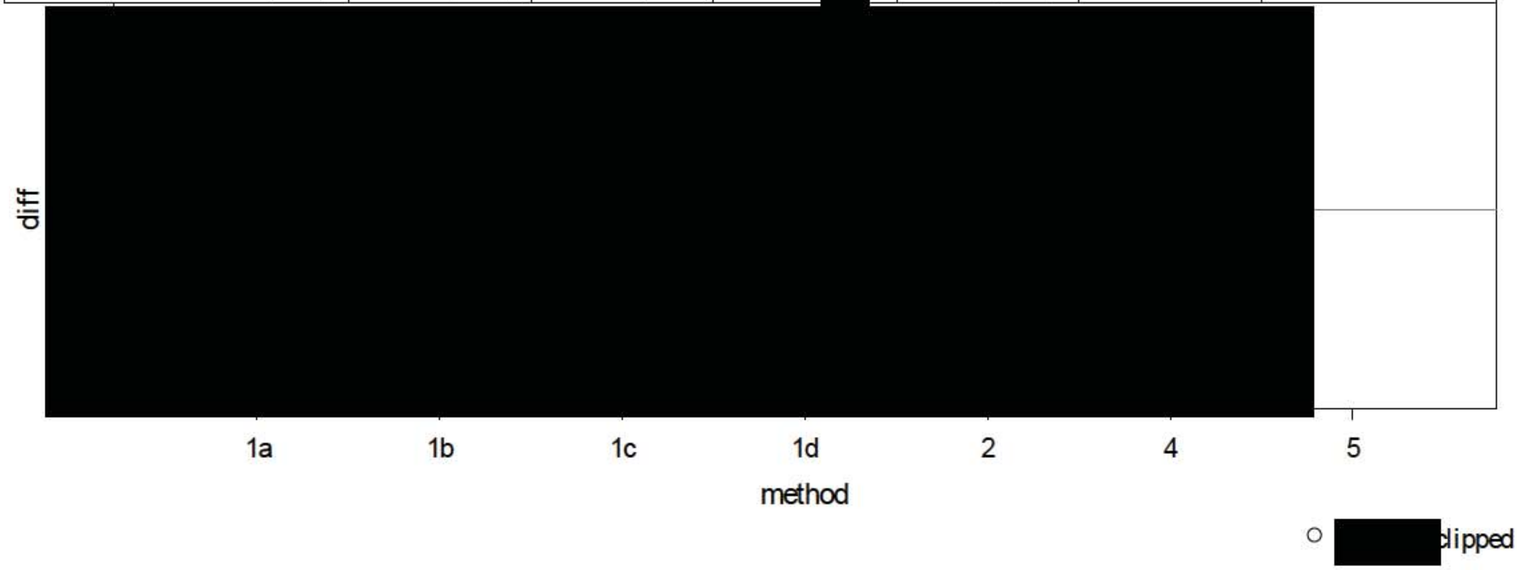
N	250	30	30	30	30	250	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<1	<15	<1	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	2.[redacted]	-19.[redacted]	1.[redacted]	-21.[redacted]	-0.[redacted]	13	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	30	<15	<15	<15	<15	20	0
Range	700	750	100	750	100	900	.
Std Dev	46.[redacted]	116.[redacted]	16.[redacted]	115.[redacted]	17.[redacted]	66.[redacted]	.



DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=706

Statistics by GQtype							
N	7300	200	200	200	200	17000	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	350	<15	<15	<1	<15	800	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1.[redacted]	-4.[redacted]	-3[redacted]	-1.[redacted]	-1.[redacted]	-0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	350	<15	<15	<15	<15	800	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	700	20	20	20	20	1600	0
Range	850	200	400	150	300	600	.
Std Dev	22.[redacted]	21.[redacted]	30[redacted]	16.[redacted]	22.[redacted]	19.[redacted]	.

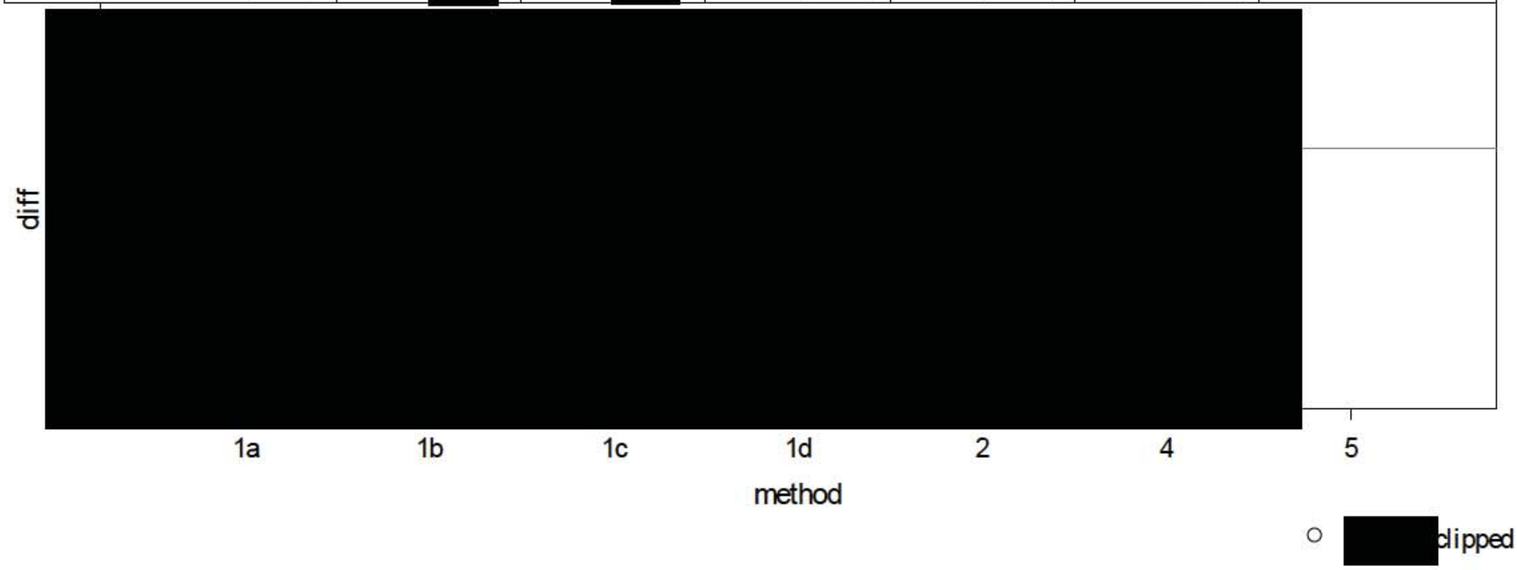


DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=801

Statistics by GQtype

N	50500	21000	21000	21000	21000	52500	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	2400	1000	1000	1000	900	1700	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-0[redacted]	-0.[redacted]	-0.[redacted]	-0[redacted]	1.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	2000	800	900	800	800	2600	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	4400	1800	1800	1800	1800	4300	0
Range	700	400	200	400	300	700	.
Std Dev	9.[redacted]	6.[redacted]	5.[redacted]	6.[redacted]	5.[redacted]	18.[redacted]	.

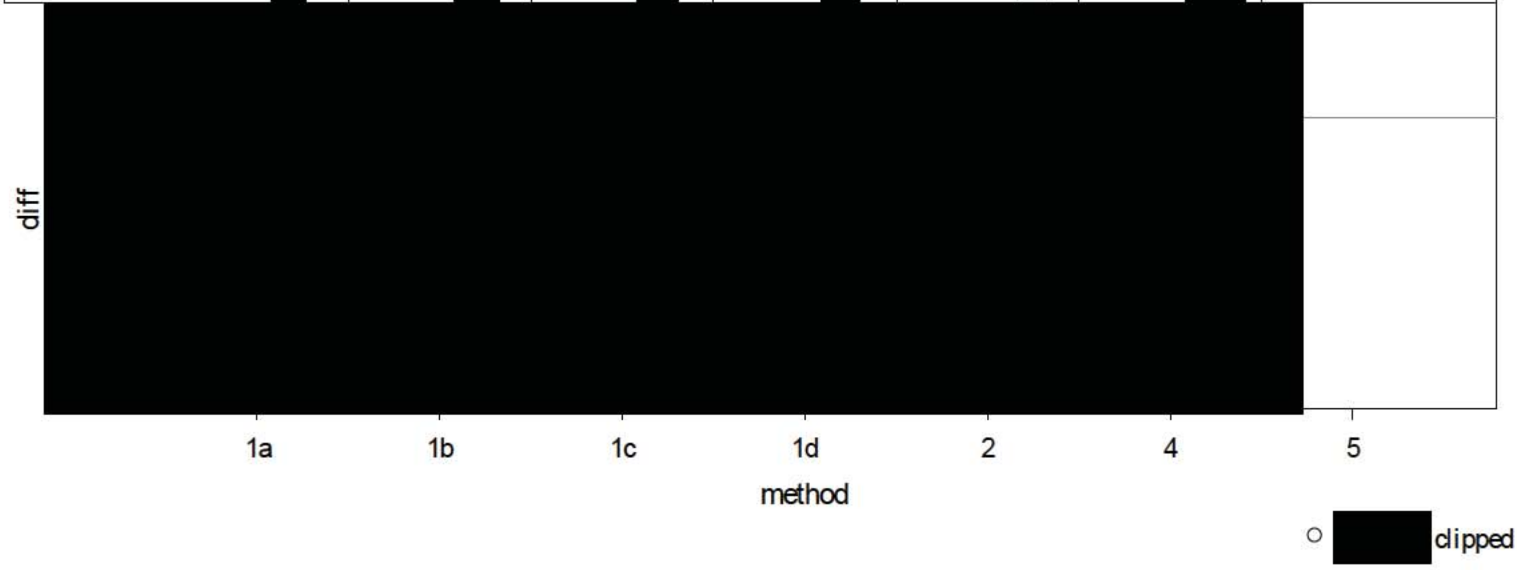


DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=802

Statistics by GQtype

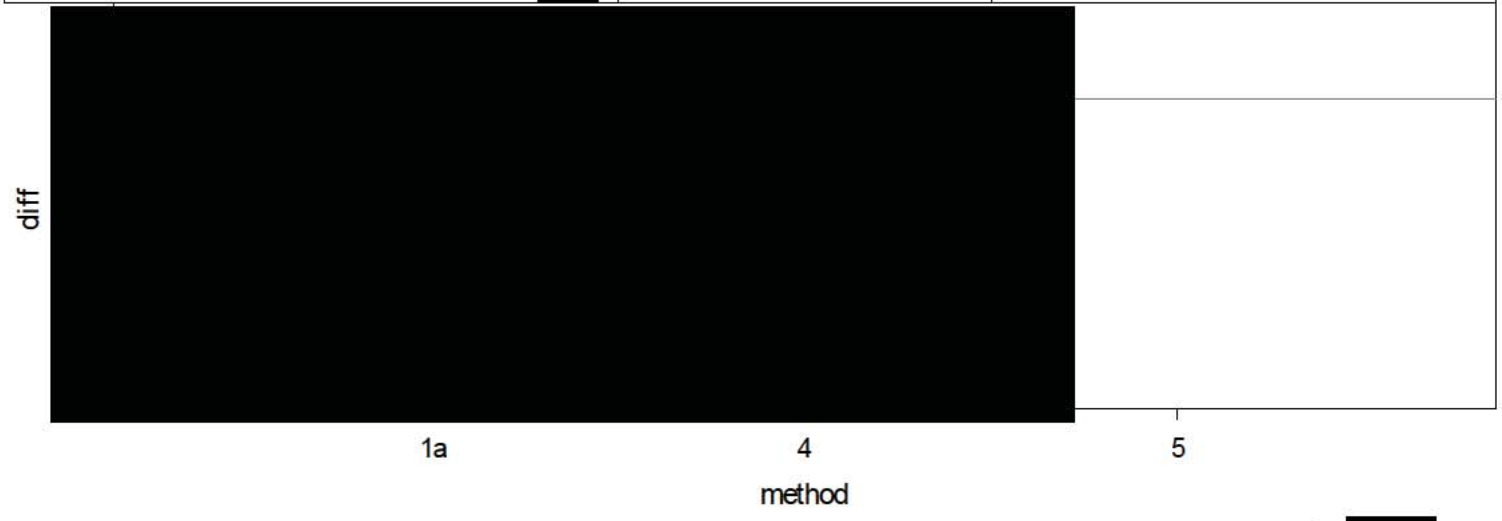
N	8000	3500	3500	3500	3500	8300	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	400	150	150	200	150	350	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	400	150	150	150	150	400	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	750	350	350	350	350	800	0
Range	1200	300	300	300	300	1200	.
Std Dev	17.[redacted]	13.[redacted]	13.[redacted]	13.[redacted]	12.[redacted]	31.[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=900

Statistics by GQtype			
N	350	350	0
Min	[redacted]	[redacted]	.
Nlow	20	0	0
Q1	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	.
Mean	-0.[redacted]	-33.[redacted]	.
Q3	[redacted]	[redacted]	.
Nhigh	20	20	0
Max	[redacted]	[redacted]	.
Nout	30	20	0
Range	150	100	.
Std Dev	10[redacted]	12[redacted]	.



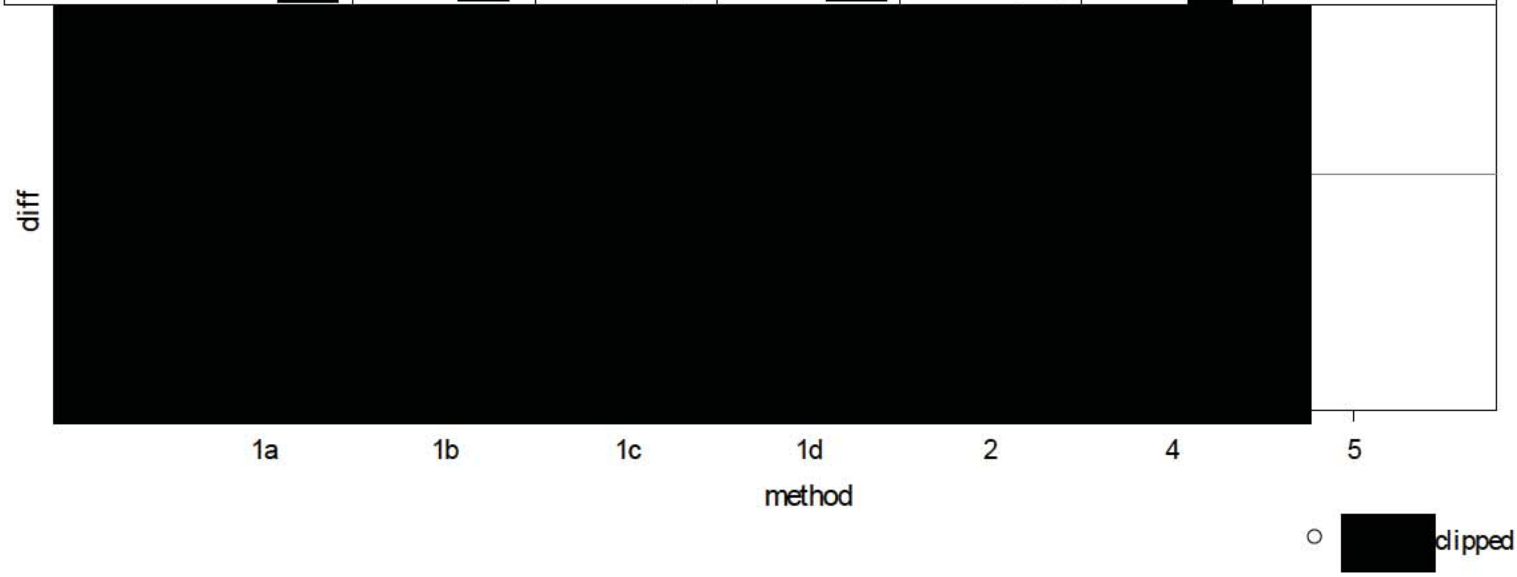
○ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=901

Statistics by GQtype

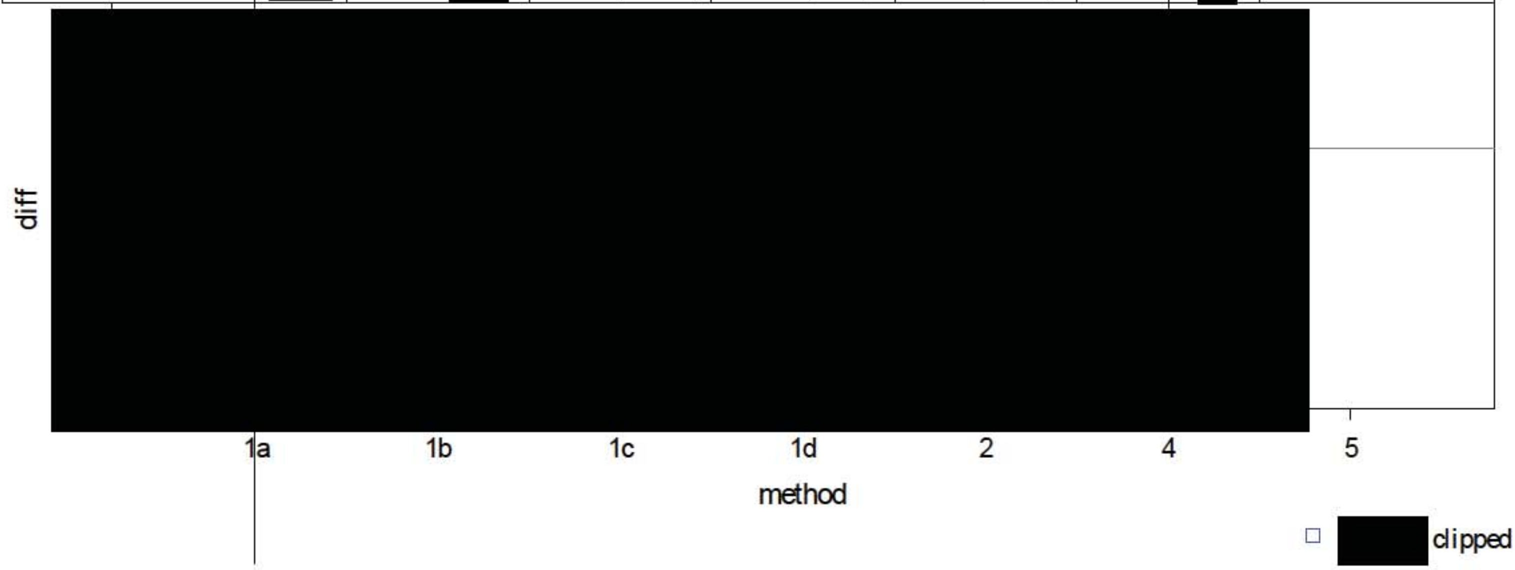
N	5300	2200	2200	2200	2200	5500	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	250	100	100	100	100	250	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	1.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	250	100	100	100	100	250	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	500	200	200	200	200	500	0
Range	1000	300	450	300	300	1100	.
Std Dev	18.[redacted]	15.[redacted]	18.[redacted]	15.[redacted]	13.[redacted]	32.[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=903

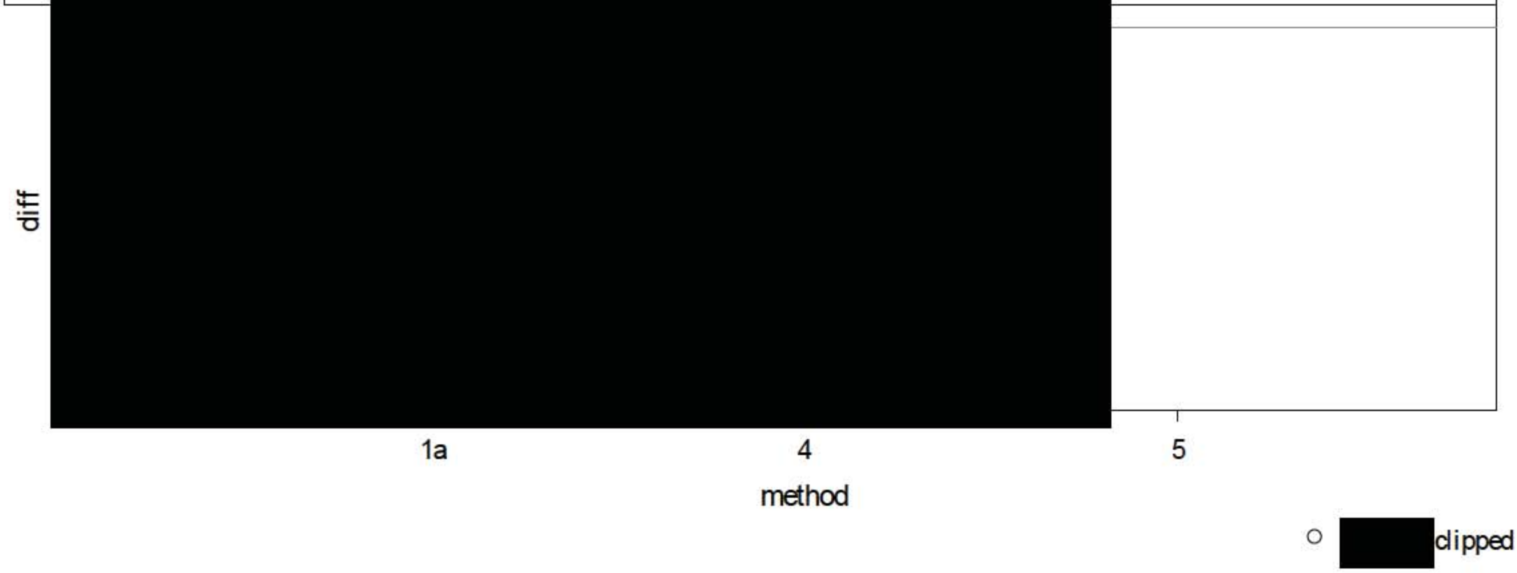
Statistics by GQtype							
N	20	<15	<15	<15	<15	20	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	0	0	0	0	0	0	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-2.[redacted]	-5.[redacted]	-4.[redacted]	0.[redacted]	7.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	0	0	0	0	0	0	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	0	0	0	0	0	0	0
Range	70	<15	20	20	<15	80	.
Std Dev	13.[redacted]	3.[redacted]	8.[redacted]	7.[redacted]	2.[redacted]	21.[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=904

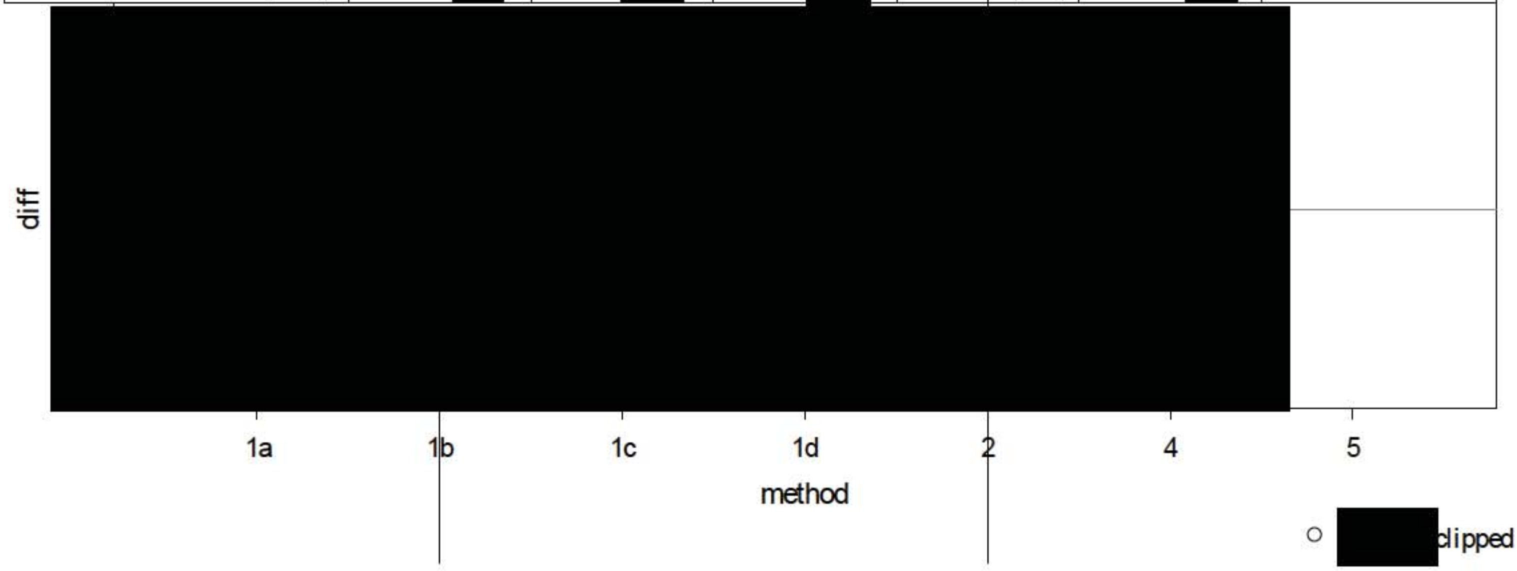
Statistics by GQtype			
N	6200	8600	0
Min	[redacted]	[redacted]	.
Nlow	300	0	0
Q1	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	.
Mean	-0.[redacted]	-30.[redacted]	.
Q3	[redacted]	[redacted]	.
Nhigh	300	450	0
Max	[redacted]	[redacted]	.
Nout	550	450	0
Range	500	1700	.
Std Dev	10[redacted]	31.[redacted]	.


[redacted]

GQTYPCUR=999

Statistics by GQtype

N	500	<15	<15	<15	<15	700	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	30	0	0	0	0	20	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-2.[redacted]	-7.[redacted]	0.[redacted]	-2.[redacted]	-2.[redacted]	12.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	30	0	0	0	0	40	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	50	0	0	0	0	50	0
Range	400	100	30	30	160	1000	.
Std Dev	24.[redacted]	21.[redacted]	5.[redacted]	8.[redacted]	35.[redacted]	49.[redacted]	.

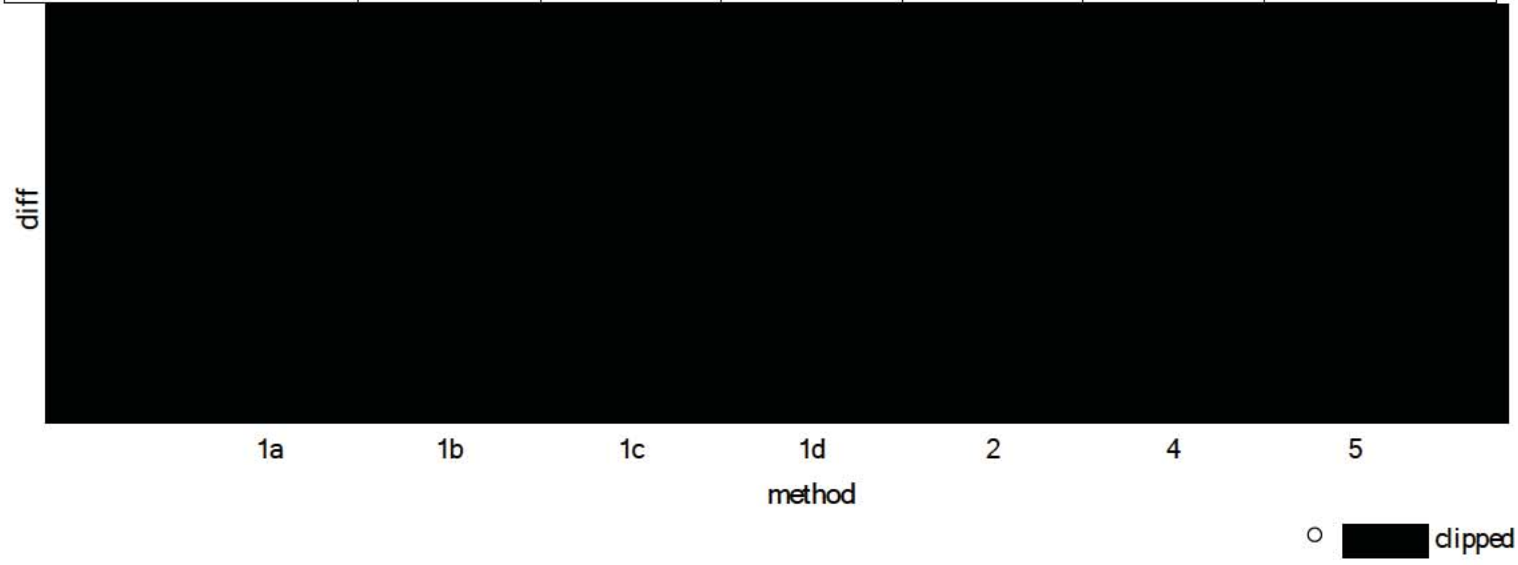

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=103

Statistics by GQtype

N	550	550	550	550	550	550	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	30	30	30	30	30	30	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-2	-7	0	-7	-1	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	30	30	30	30	30	30	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	50	50	50	50	50	50	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	162	172	171	172	139	278	.

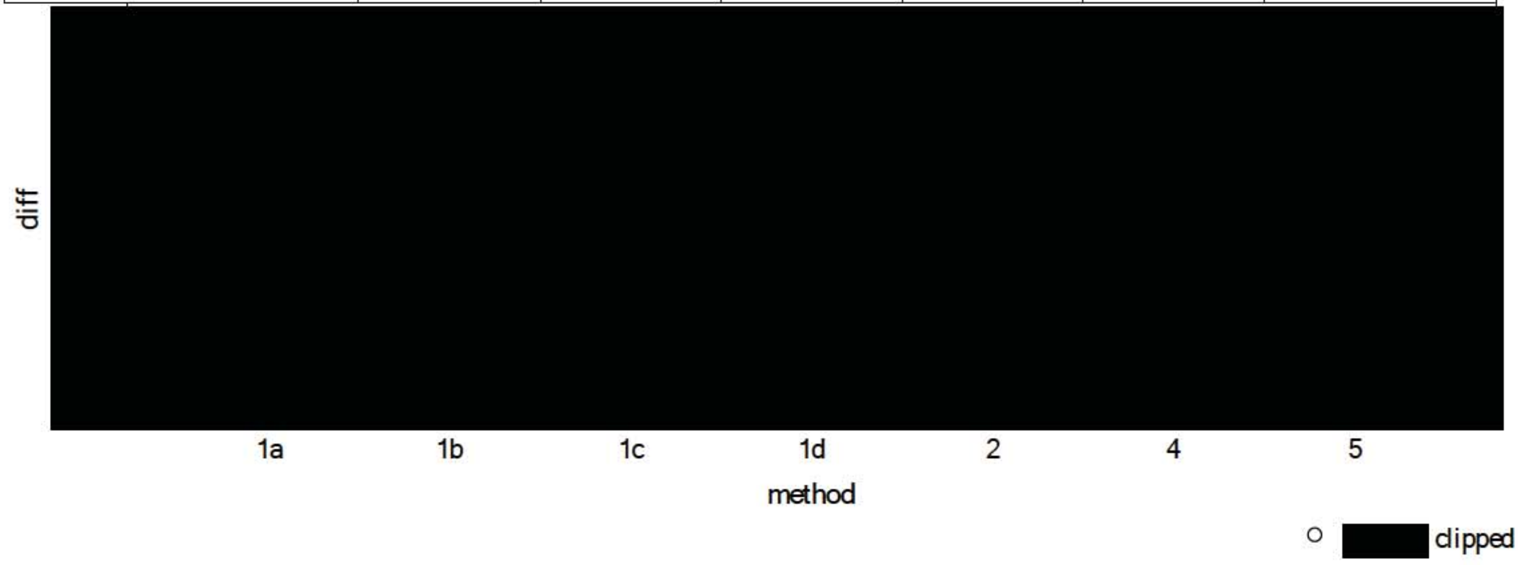

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=104

Statistics by GQtype

N	2000	2000	2000	2000	2000	2000	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	100	100	100	100	100	100	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1	-1	-1	-3	-1	6	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	100	100	100	100	100	100	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	200	200	200	200	200	200	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	118	137	114	138	145	274	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=105

Statistics by GQtype							
N	350	350	350	350	350	350	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	20	20	20	20	20	20	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-4	-2	2	-3	-1	2	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	20	20	20	20	20	20	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	30	30	30	30	30	30	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	65	62	64	66	47	113	.

diff

[redacted]

1a 1b 1c 1d 2 4 5

method

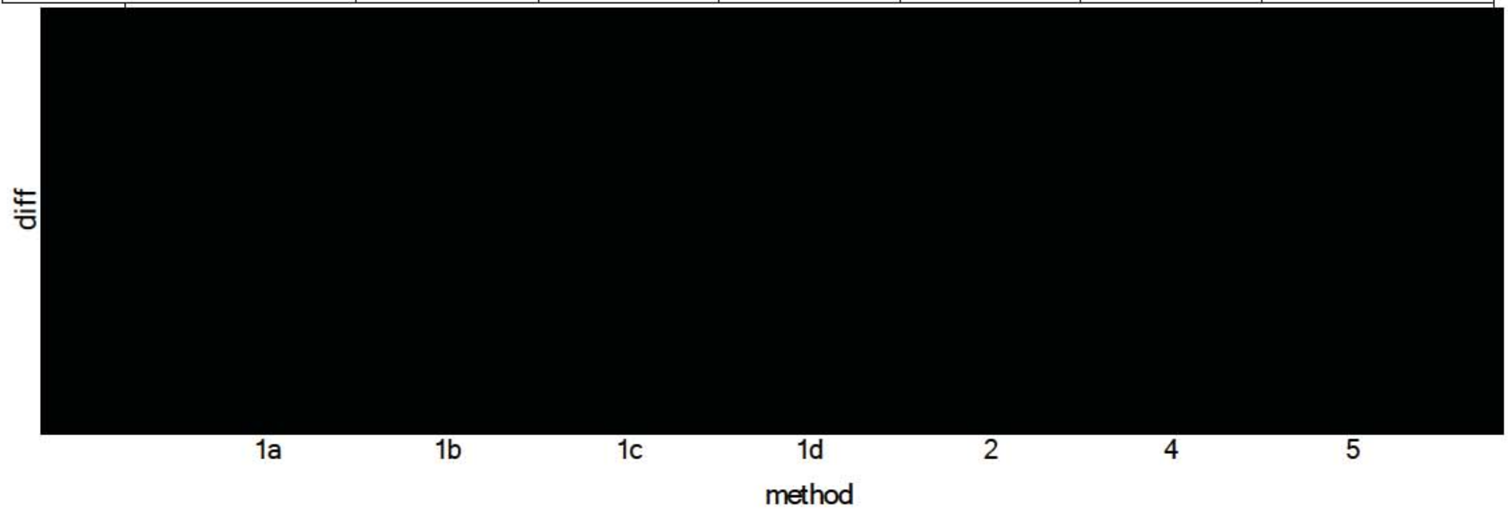
○ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=106

Statistics by GQtype							
N	<15	<15	<15	<15	<15	<15	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	0	0	0	0	0	0	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1	-1	24	17	-1	-23	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	0	0	0	0	0	0	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	0	0	0	0	0	0	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	51	57	56	45	73	149	.

[redacted]

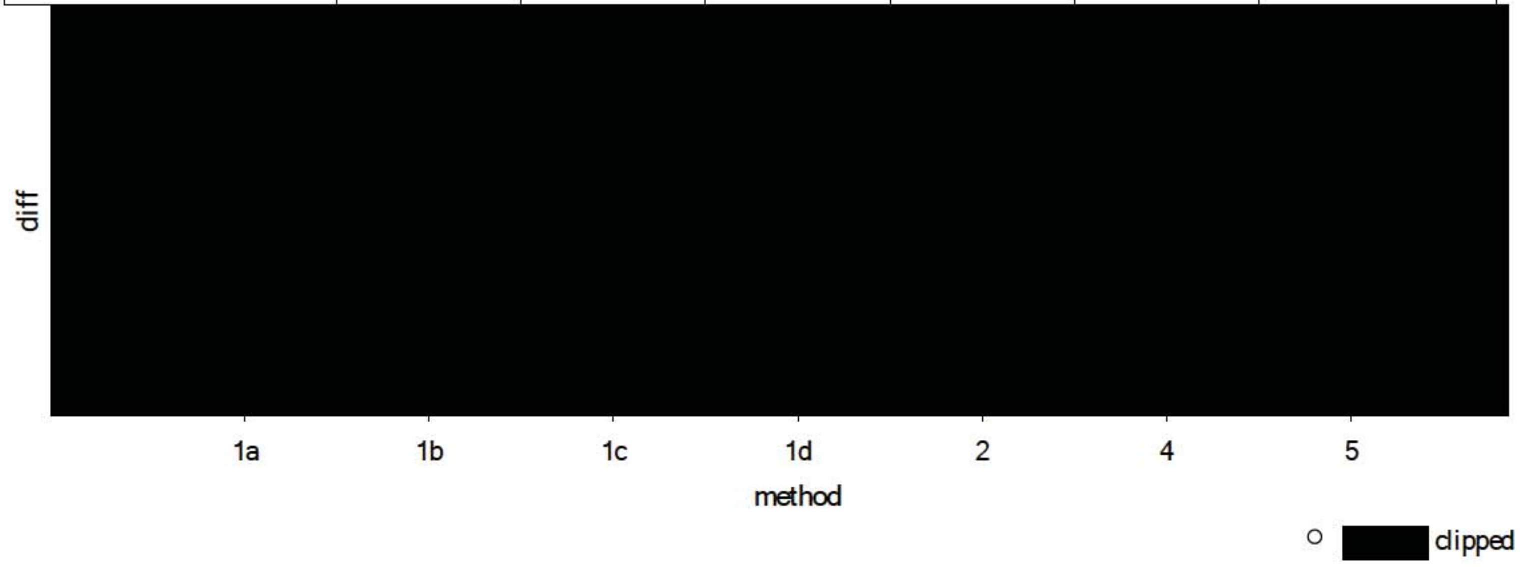


□ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=201

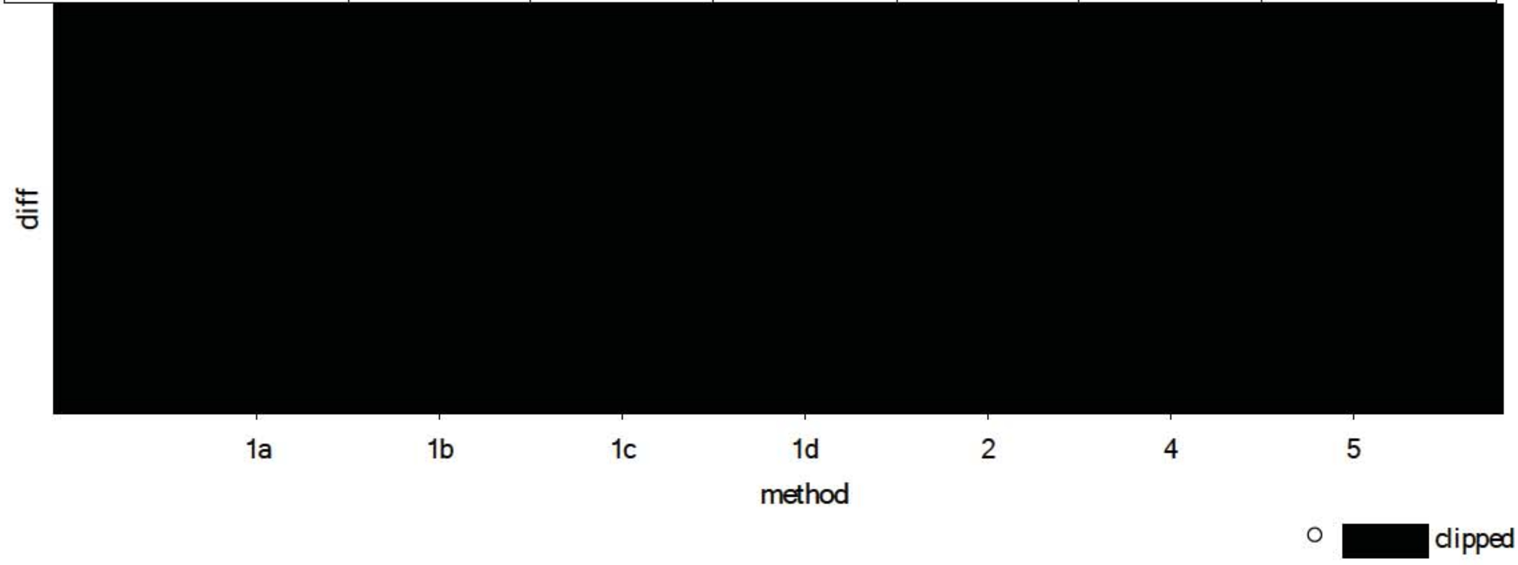
Statistics by GQtype							
N	1400	1400	1400	1400	1400	1400	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	70	60	70	60	70	60	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0	0	-1	-1	0	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	60	70	70	60	70	70	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	150	150	150	100	150	150	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	9	9	8	9	8	13	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=202

Statistics by GQtype							
N	850	850	850	850	850	850	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	40	40	40	40	40	40	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0	–1	0	-0	-1	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	40	40	40	40	40	40	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	80	80	80	80	80	80	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	11	13	14	13	12	24	.

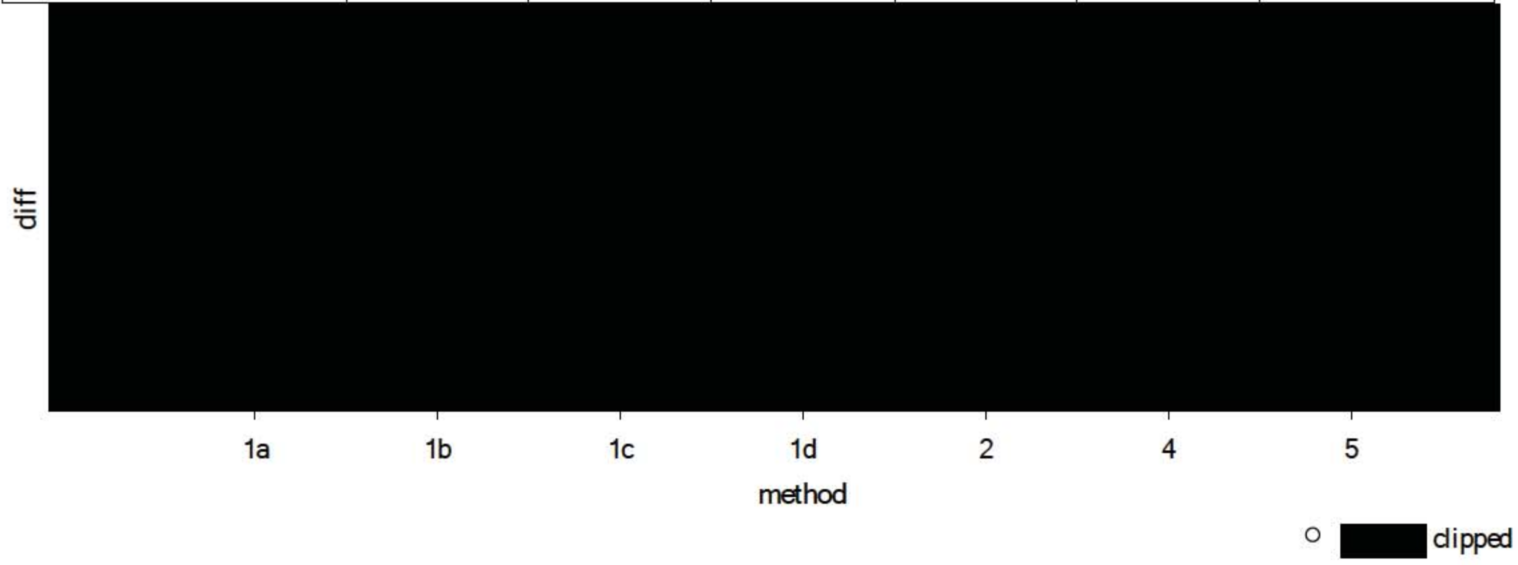

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=203

Statistics by GQtype

N	550	550	550	550	550	550	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	30	30	30	30	30	30	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-2	–3	-2	–3	-1	0	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	30	30	30	30	30	30	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	60	60	60	50	60	50	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	29	33	29	31	19	24	.

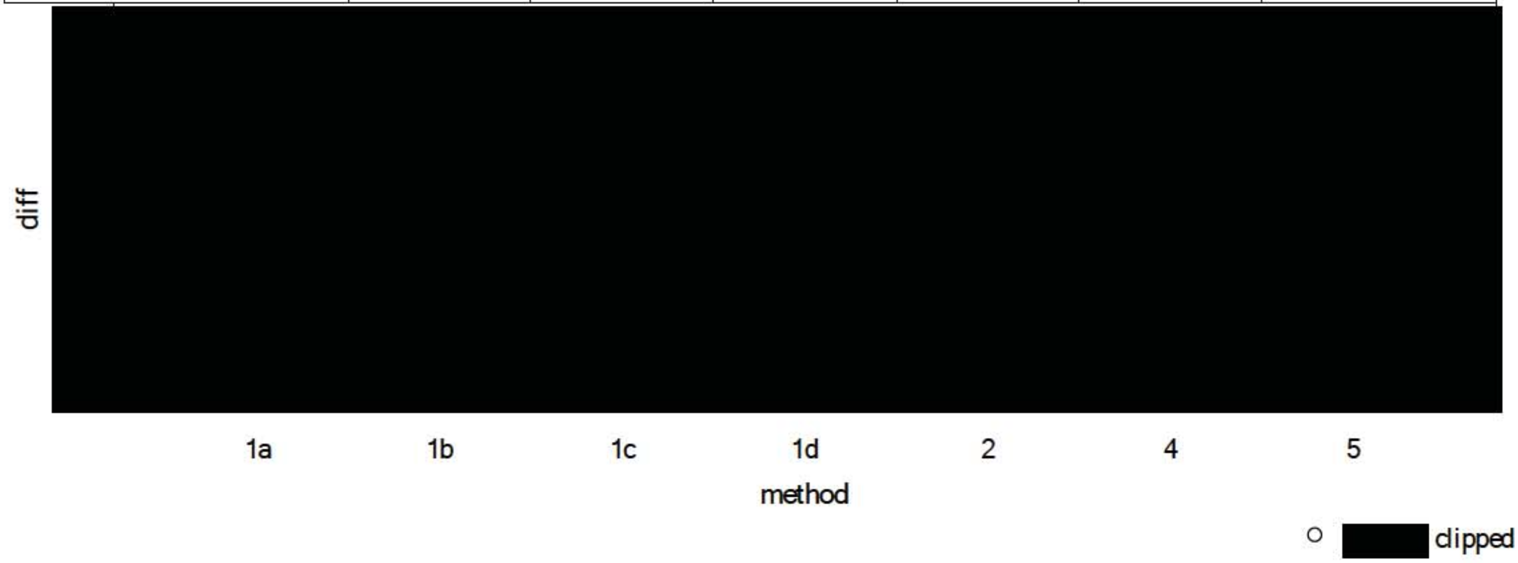

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=301

Statistics by GQtype

N	14500	14500	14500	14500	14500	14500	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	700	700	700	700	700	700	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0	0	-0	0	–1	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	700	700	700	700	700	700	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	1400	1400	1400	1400	1400	1400	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	25	28	28	28	25	50	.



DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=401

Statistics by GQtype							
N	250	250	250	250	250	250	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-5	-4	0	0	-1	3	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	20	20	20	20	20	20	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	82	83	63	79	46	101	.

diff

[redacted]

1a 1b 1c 1d 2 4 5

method

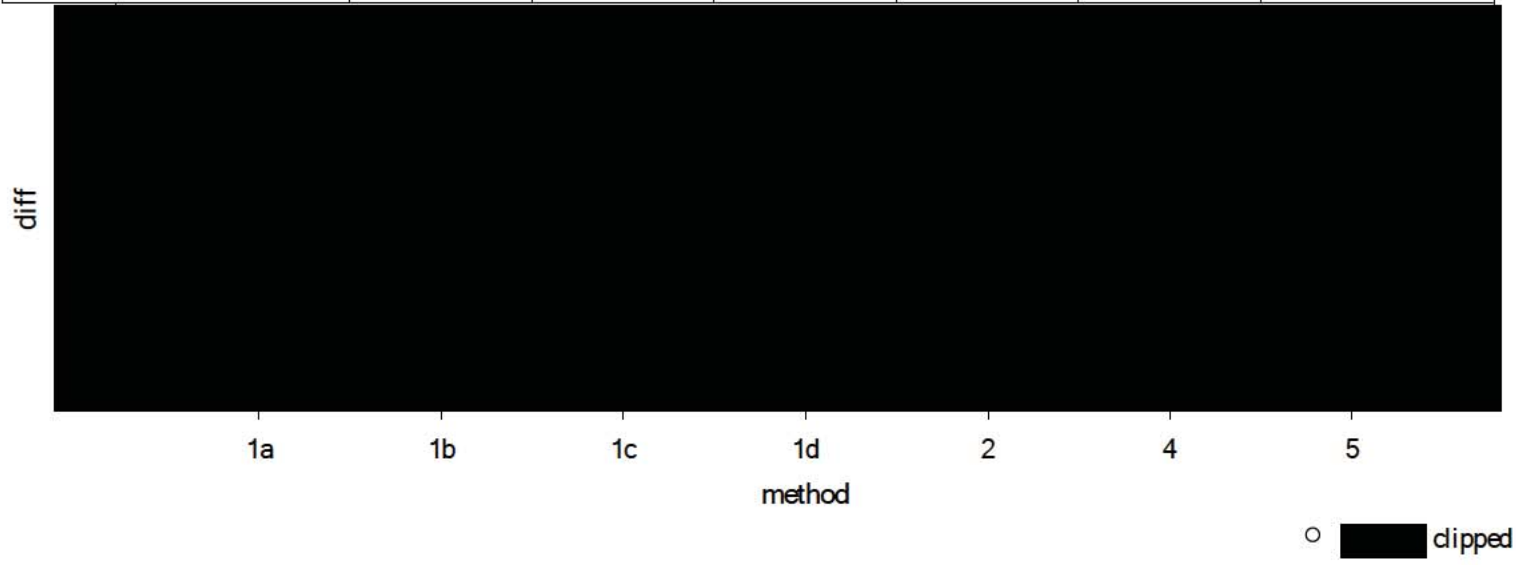
○ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=402

Statistics by GQtype

N	70	70	70	70	70	70	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-5	–3	-7	-4	0	-6	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	<15	<15	<15	<15	<15	<15	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	30	26	33	43	31	74	.

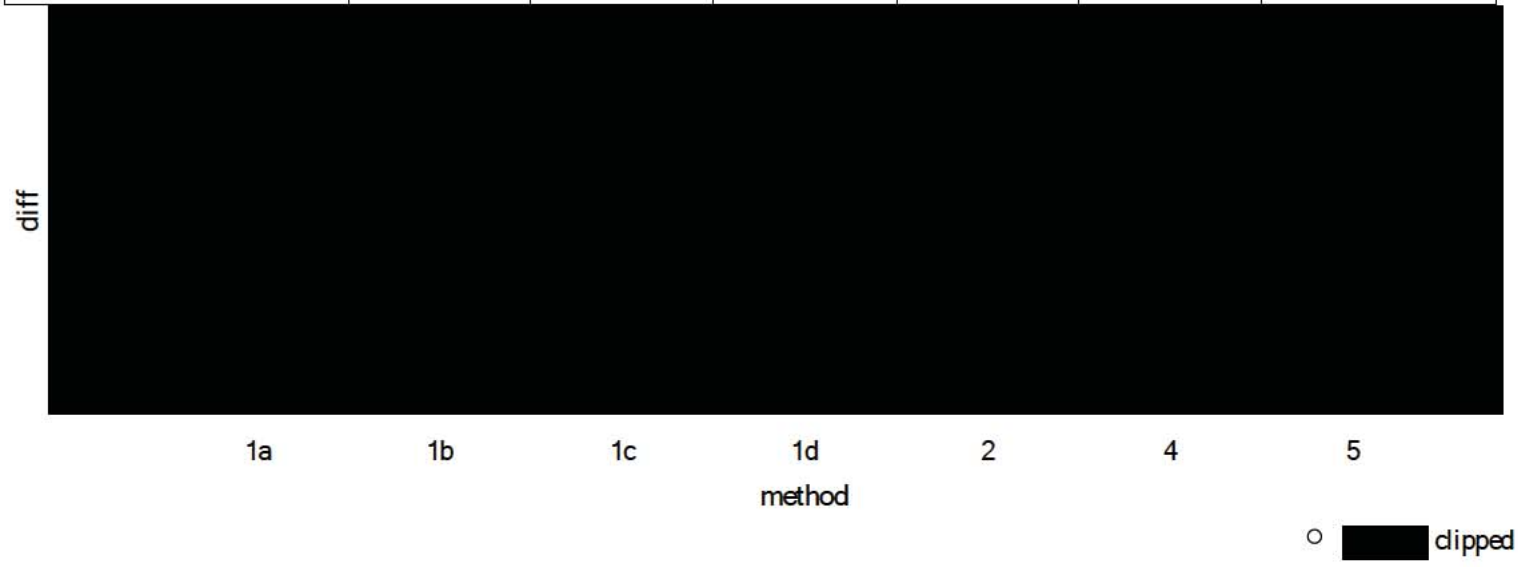

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=403

Statistics by GQtype

N	200	200	200	200	200	200	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0	0	-1	0	-1	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	20	20	20	20	20	20	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	11	9	12	9	11	25	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=404

Statistics by GQtype							
N	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.

diff

[redacted]

1a 1b 1c 1d 2 4 5

method

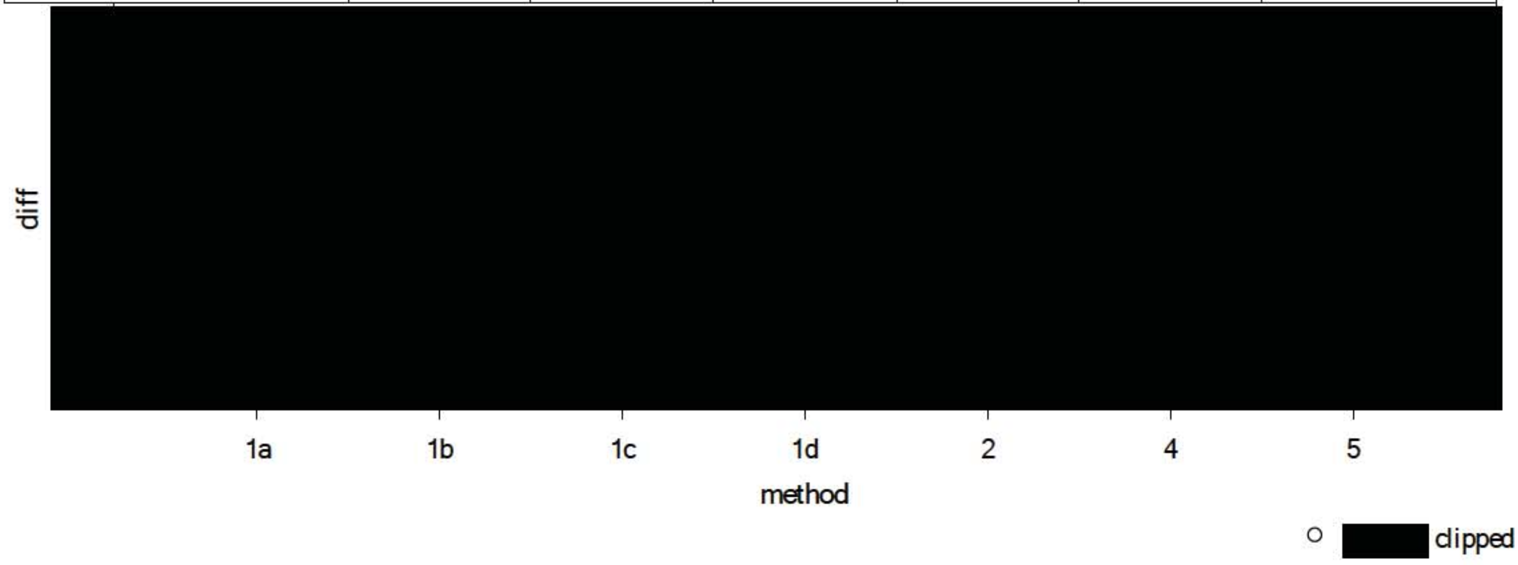
□ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=405

Statistics by GQtype

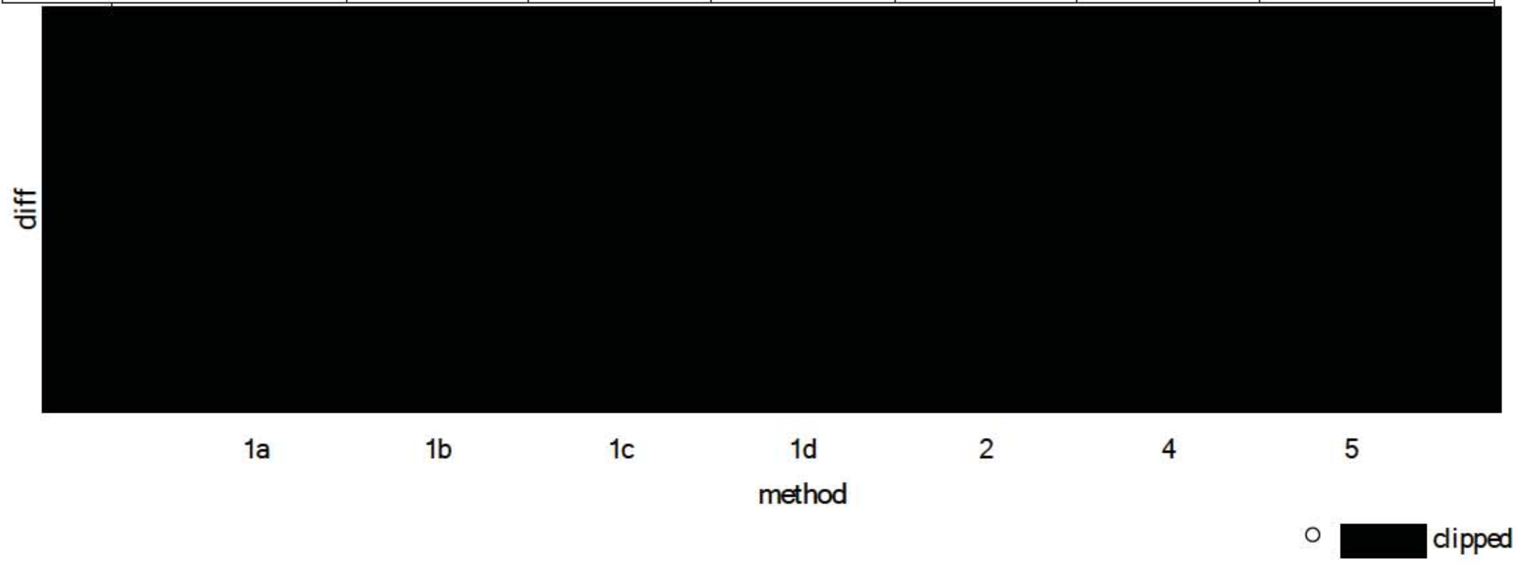
N	250	250	250	250	250	250	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-3	-4	0	-3	-1	-1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	20	20	20	20	20	20	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	29	31	26	30	22	49	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=501

Statistics by GQtype							
N	13000	13000	13000	13000	13000	13000	13000
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nlow	650	650	650	650	650	650	650
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	0	-1	0	-2	0	-2	0
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nhigh	650	650	650	650	650	650	650
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nout	1300	1300	1300	1300	1300	1300	1300
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Std Dev	34	42	39	43	35	128	46


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=601

Statistics by GQtype

N	800	800	800	800	800	800	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	40	40	40	40	40	40	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-4	-4	-5	-2	-1	-4	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	40	40	40	40	40	40	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	80	80	80	80	80	80	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	94	96	102	94	82	107	.

diff

[redacted]

1a 1b 1c 1d 2 4 5

method

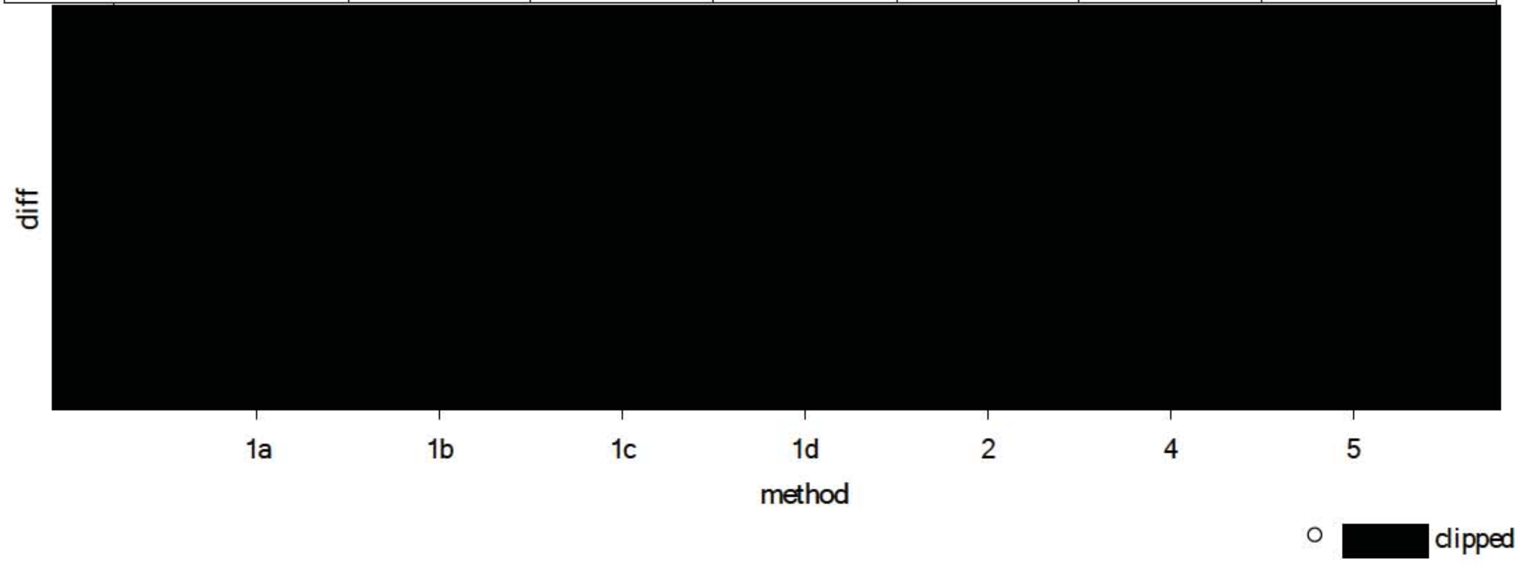
○ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=701

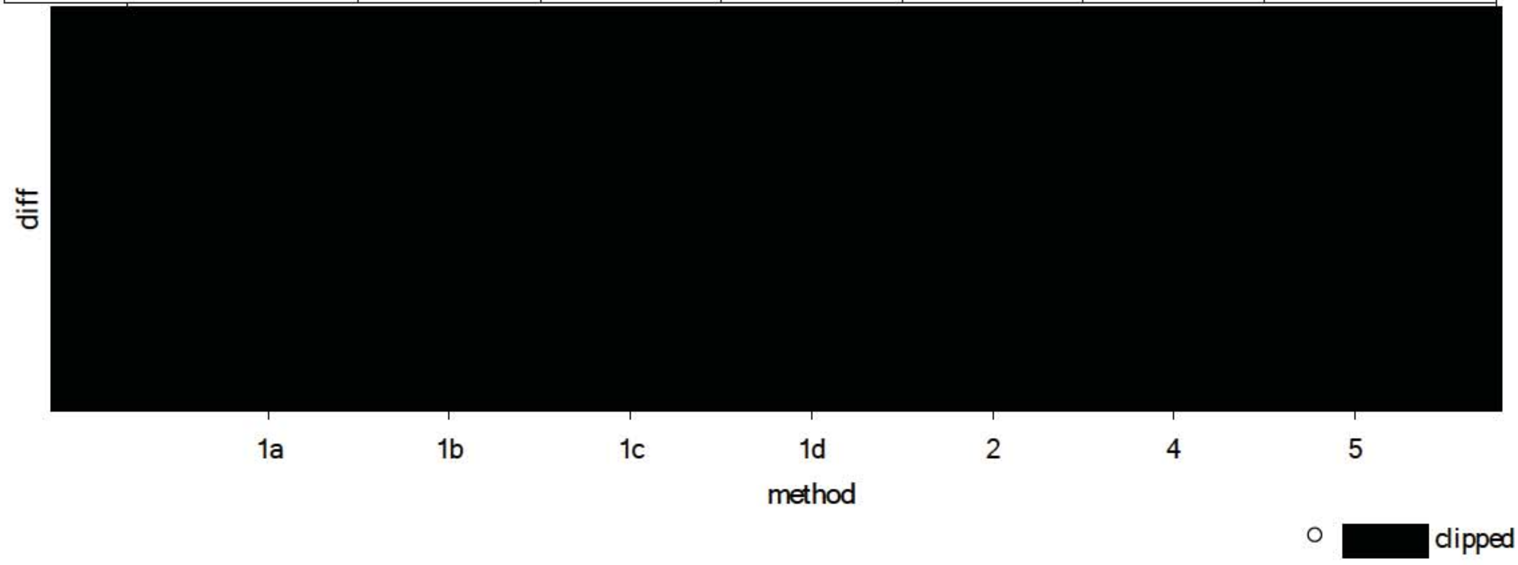
Statistics by GQtype

N	1600	1600	1600	1600	1600	1600	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	80	80	80	80	80	80	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1	-1	0	-3	0	5	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	80	80	80	80	80	80	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	150	150	150	150	150	150	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	37	36	43	37	33	59	.


[redacted]

GQTYPCUR=702

Statistics by GQtype							
N	700	700	700	700	700	700	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	40	40	40	40	40	30	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-3	-7	-2	-13	0	3	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	40	40	40	40	40	40	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	70	70	70	70	70	70	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	63	59	82	62	62	92	.

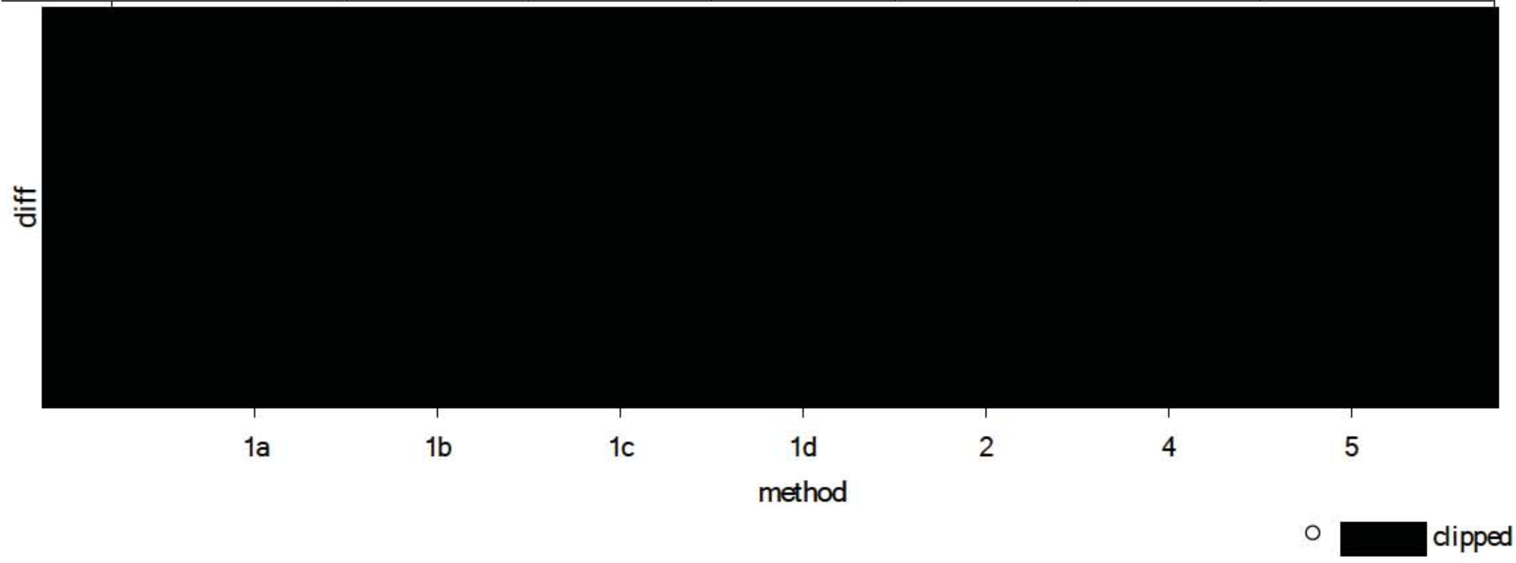

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=704

Statistics by GQtype

N	30	30	30	30	30	30	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	2	-20	2	-21	0	6	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	<15	<15	<15	<15	<15	<15	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	21	116	17	116	18	24	.

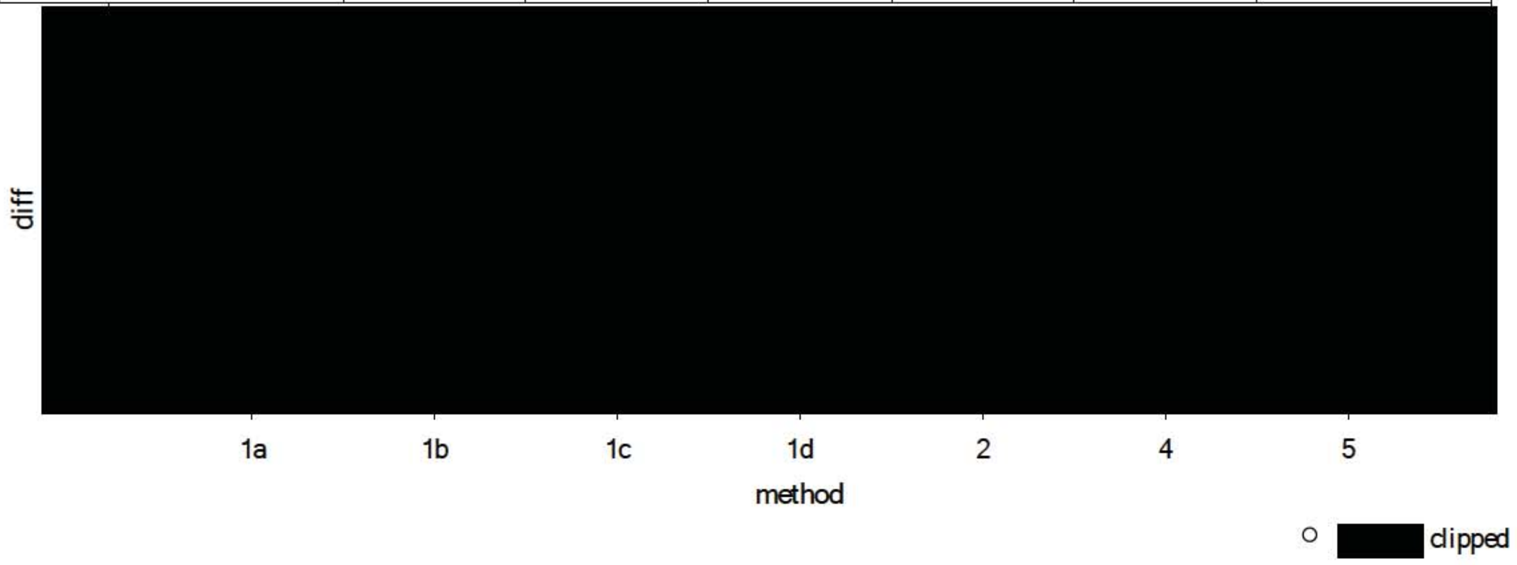

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=706

Statistics by GQtype

N	200	200	200	200	200	200	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1	-5	-3	-1	-1	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	20	20	20	20	20	20	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	18	22	31	17	22	19	.

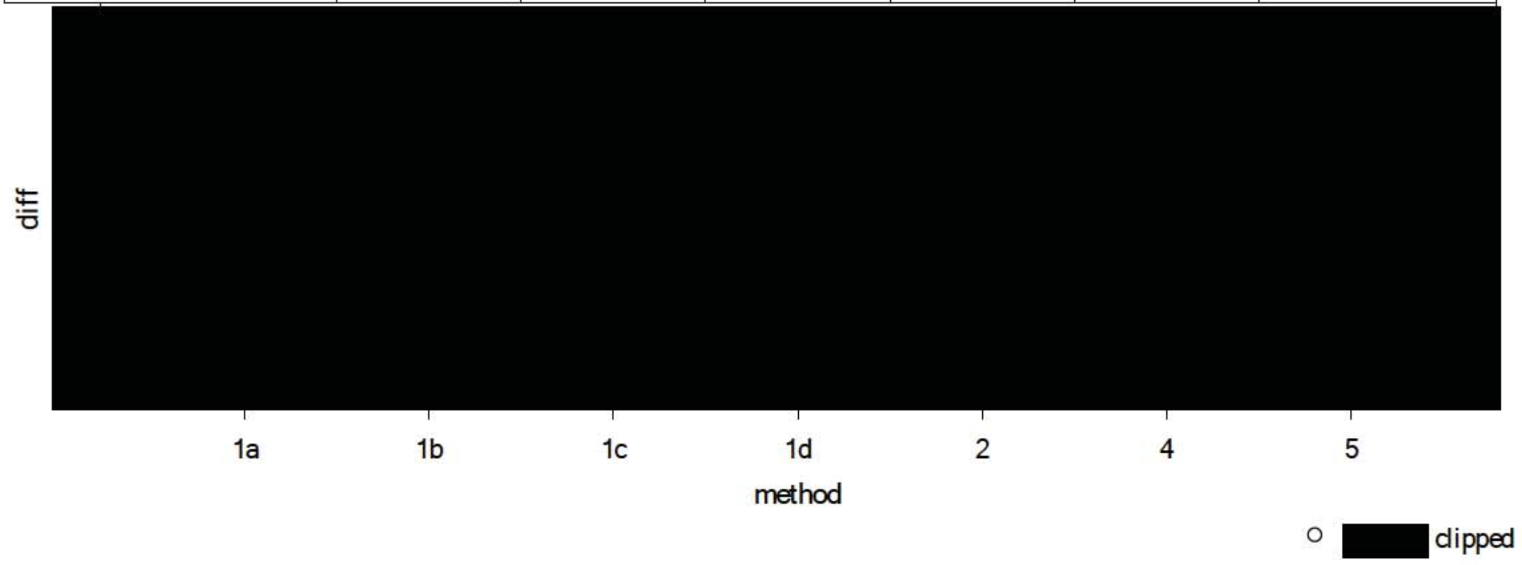

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=801

Statistics by GQtype

N	21000	21000	21000	21000	21000	21000	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	850	1000	950	1000	900	850	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0	0	0	0	-1	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	950	800	900	750	850	950	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	1800	1800	1800	1800	1800	1800	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	5	6	6	6	6	12	.

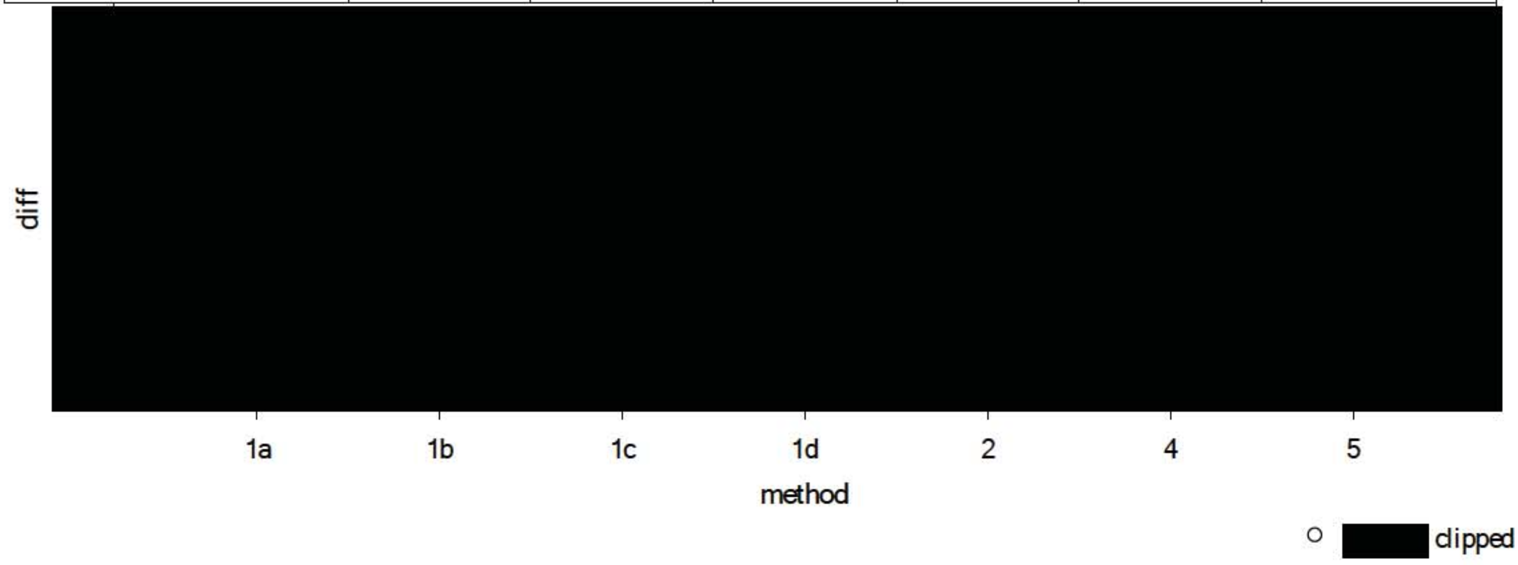

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=802

Statistics by GQtype

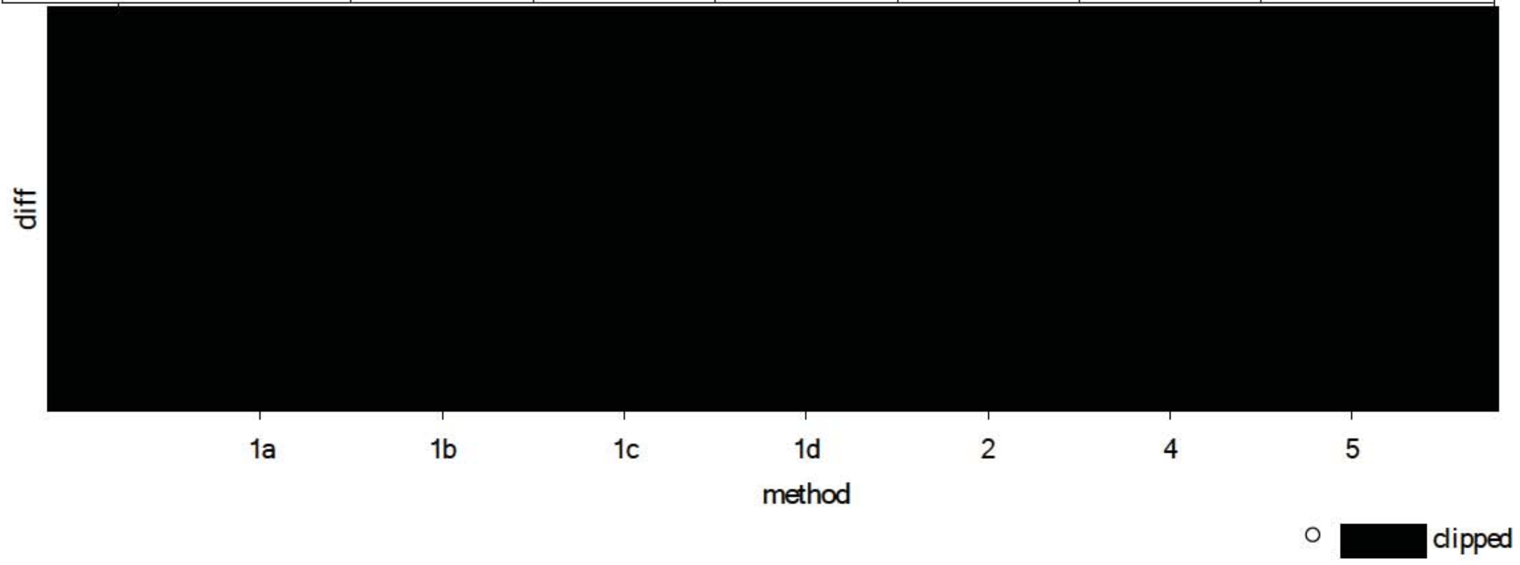
N	3500	3500	3500	3500	3500	3500	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	150	150	150	200	150	150	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1	-1	0	-1	-1	2	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	150	150	150	150	150	150	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	350	350	350	350	350	350	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	12	13	13	13	12	29	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=901

Statistics by GQtype							
N	2200	2200	2200	2200	2200	2200	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	100	100	100	100	100	100	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1	-1	-1	–1	0	3	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	100	100	100	100	100	100	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	200	200	200	200	200	200	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	15	15	19	15	14	27	.

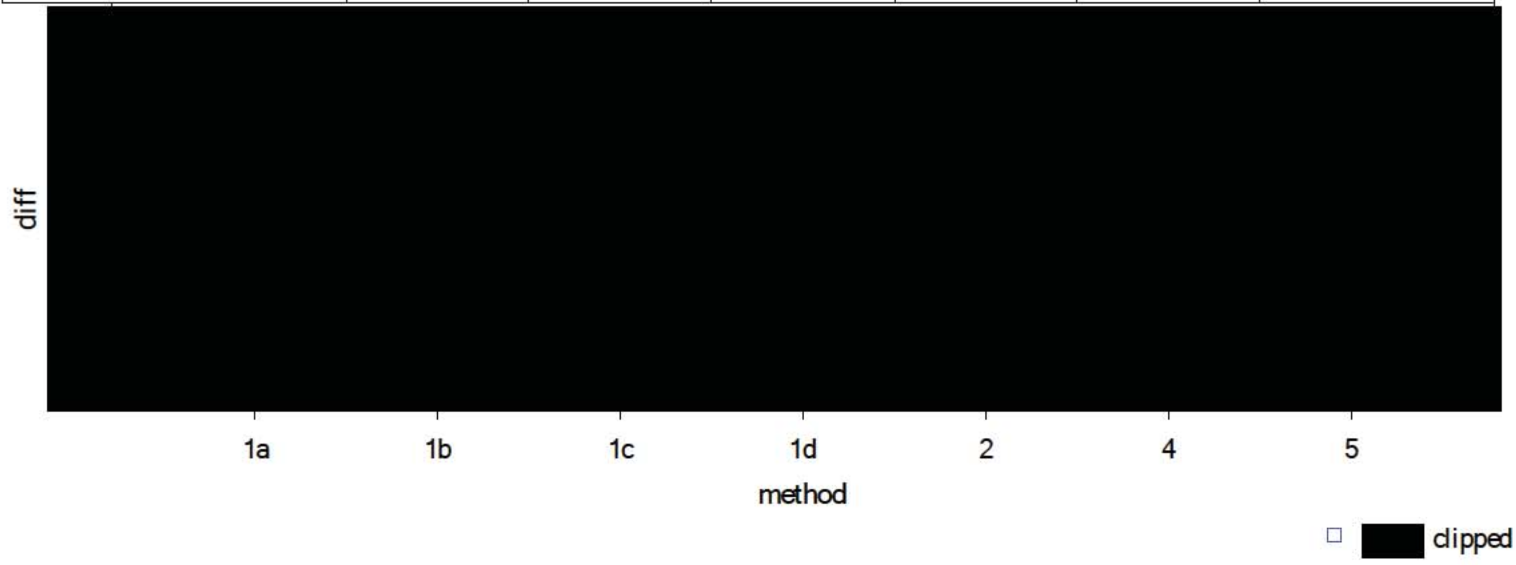

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=903

Statistics by GQtype

N	<15	<15	<15	<15	<15	<15	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	0	0	0	0	0	0	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	6	-2	-6	-4	0	17	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	0	0	0	0	0	0	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	0	0	0	0	0	0	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	19	3	9	7	2	36	.

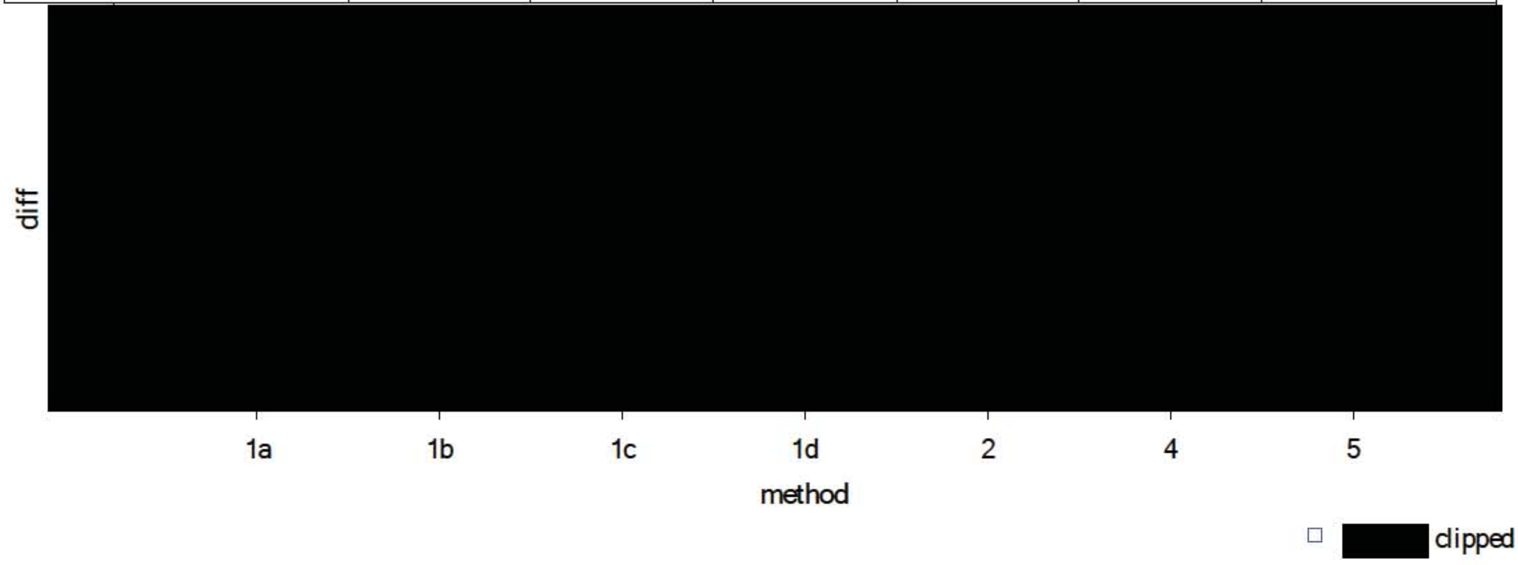

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=999

Statistics by GQtype

N	<15	<15	<15	<15	<15	<15	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	0	0	0	0	0	0	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-17	-8	0	-3	-2	14	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	0	0	0	0	0	0	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	0	0	0	0	0	0	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	51	22	6	8	36	34	.


[redacted]

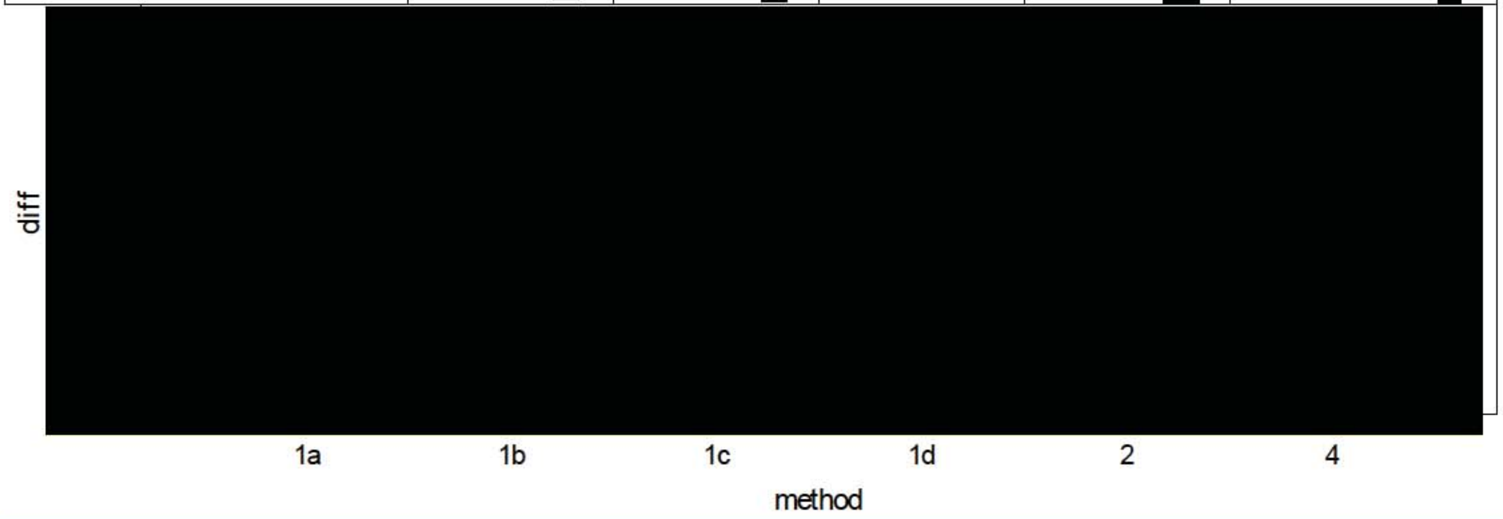
DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=103

Overall Statistics					
Min	[redacted]	Mean	3.[redacted]	Max	[redacted]
Pooled Std Dev	264.[redacted]				

Statistics by GQtype

N	550	550	550	550	550	550
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-8.[redacted]	28.[redacted]	-8.[redacted]	-0.[redacted]	12.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	3100	3500	6500	3500	4600	6200
Std Dev	195.[redacted]	219.[redacted]	313.[redacted]	218.[redacted]	234[redacted]	363.[redacted]

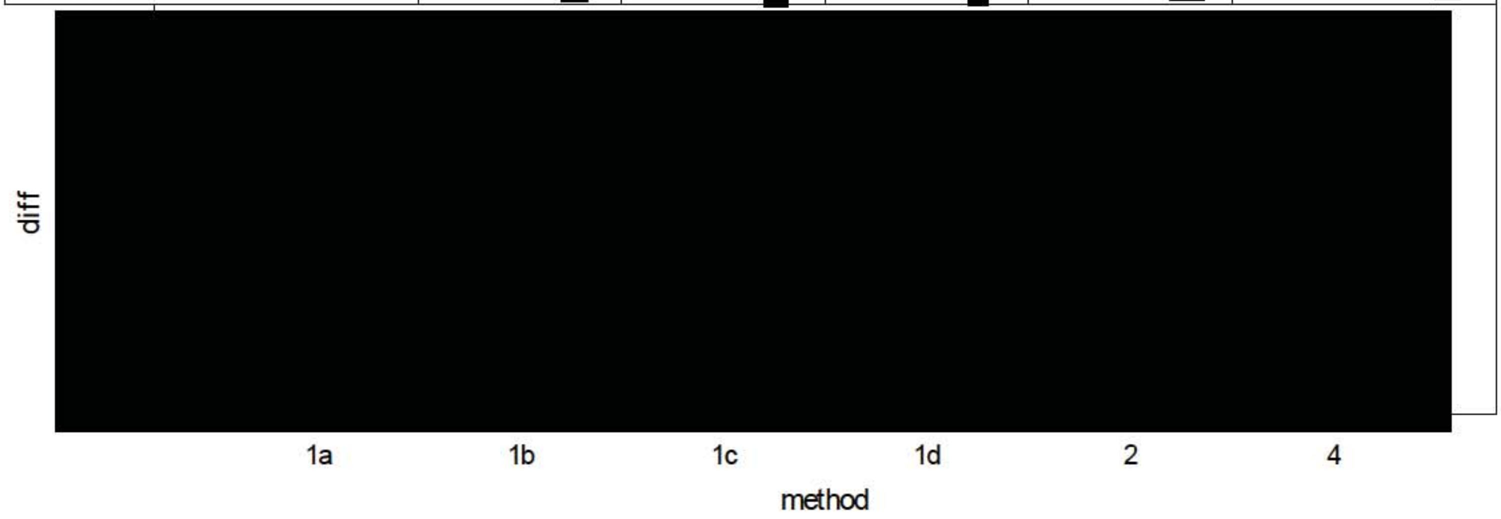

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=104

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	205.[redacted]				

Statistics by GQtype						
N	2100	2100	2100	2100	2100	2100
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-1.[redacted]	-8.[redacted]	11.[redacted]	-12.[redacted]	-0.[redacted]	5.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	3000	8200	3100	10500	3800	4200
Std Dev	139.[redacted]	220.[redacted]	155.[redacted]	250.[redacted]	137.[redacted]	283.[redacted]

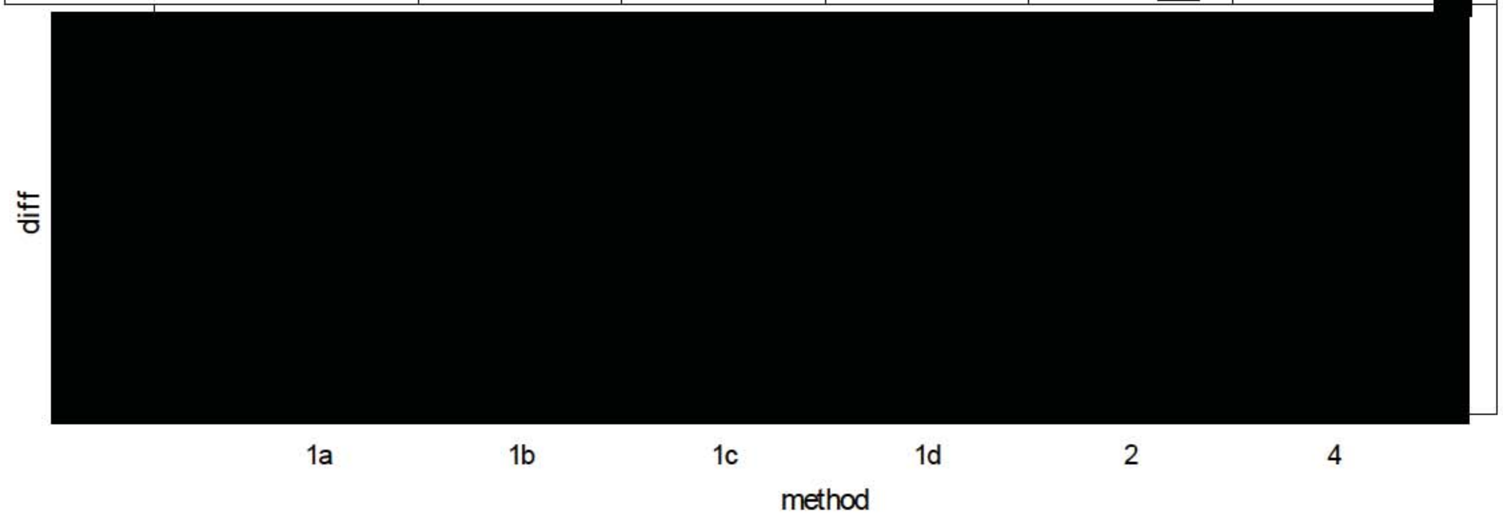

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=105

Overall Statistics					
Min	[redacted]	Mean	-74.[redacted]	Max	[redacted]
Pooled Std Dev	2398				

Statistics by GQtype						
N	350	350	350	350	350	350
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-232.[redacted]	6.[redacted]	-227.[redacted]	-0.[redacted]	4.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	1100	77000	1300	74500	800	1800
Std Dev	74.[redacted]	4223	90.[redacted]	4079	62.[redacted]	137.[redacted]



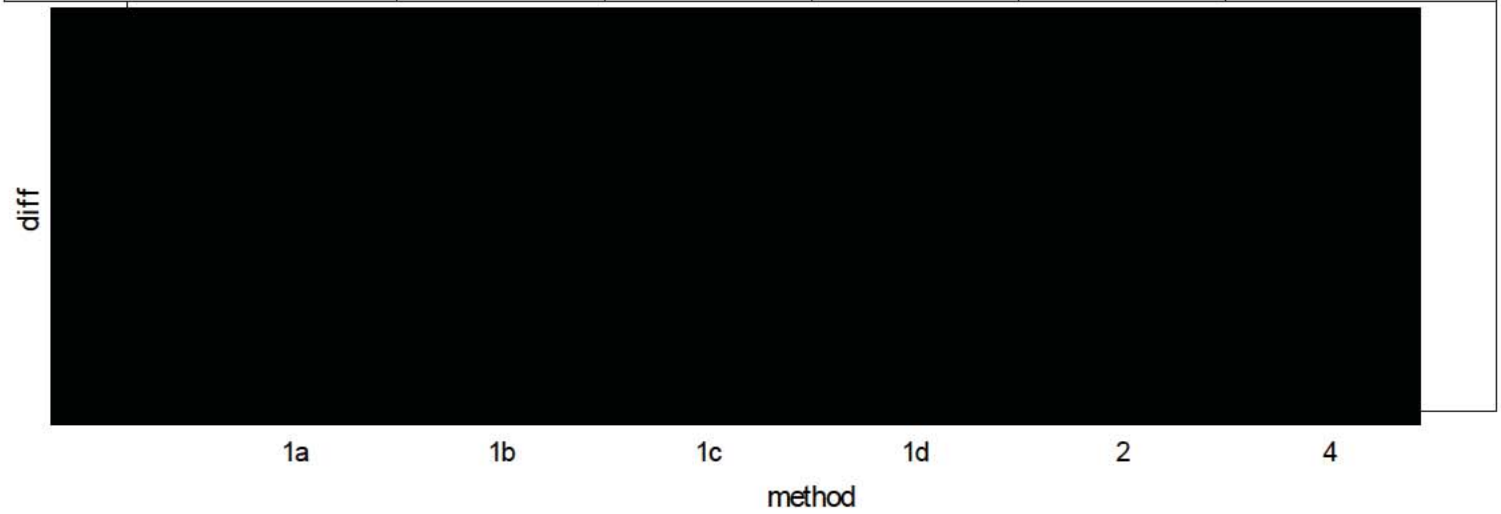
[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=106

Overall Statistics					
Min	[redacted]	Mean	8.[redacted]	Max	[redacted]
Pooled Std Dev	79.[redacted]				

Statistics by GQtype						
N	<15	<15	<15	<15	<15	<15
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	2	1	26	17	5	-2
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	150	150	150	100	200	350
Std Dev	51	56	55	44	68	148


[redacted]

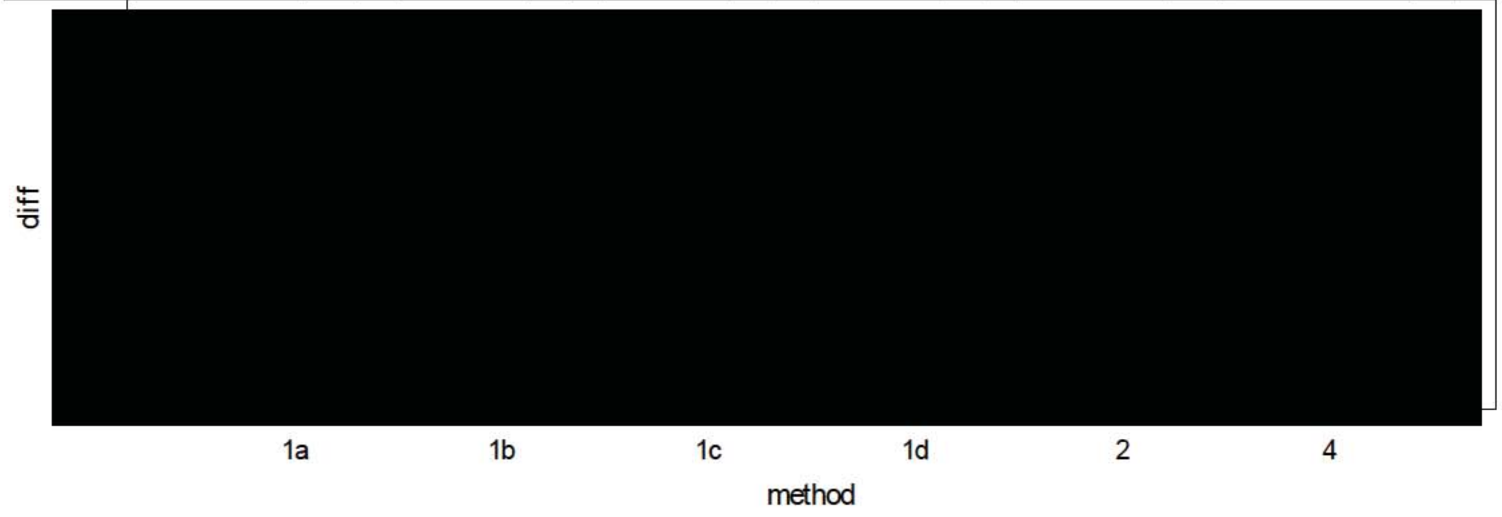
DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=201

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	11.[redacted]				

Statistics by GQtype

N	1500	1500	1500	1500	1500	1500
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	200	200	450	200	200	250
Std Dev	9.[redacted]	10.[redacted]	13.[redacted]	10.[redacted]	10.[redacted]	15.[redacted]

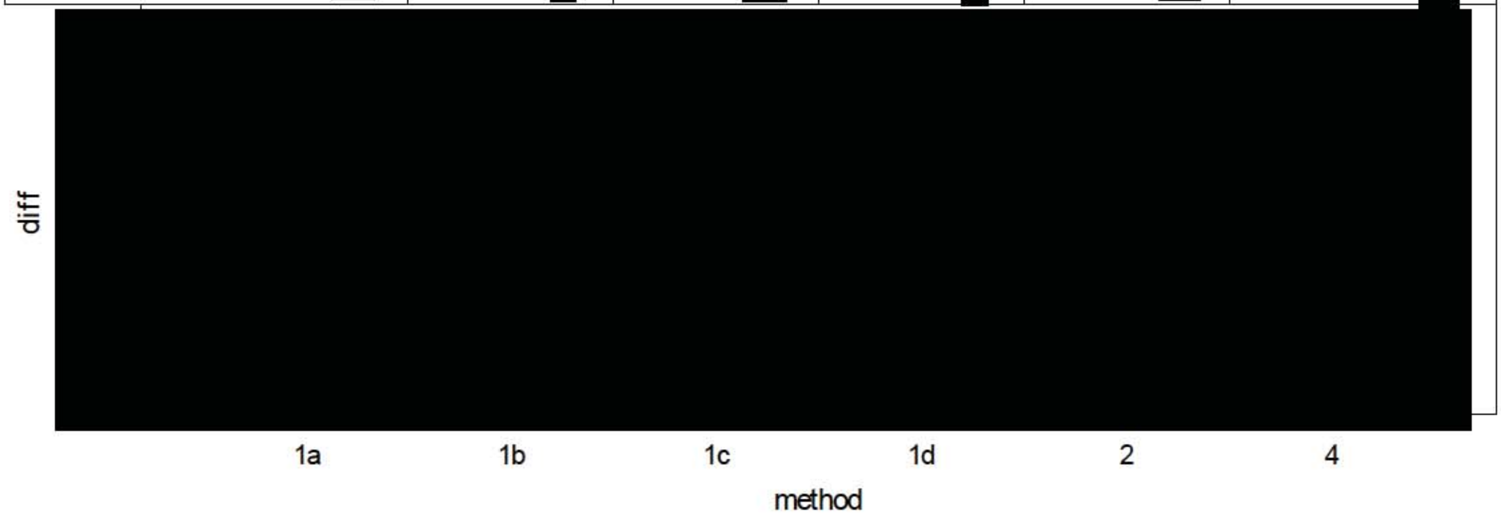

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=202

Overall Statistics					
Min	[redacted]	Mean	-2.[redacted]	Max	[redacted]
Pooled Std Dev	92.[redacted]				

Statistics by GQtype						
N	850	850	850	850	850	850
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-7.[redacted]	-0.[redacted]	-6.[redacted]	-0.[redacted]	1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	200	4700	300	4700	500	300
Std Dev	12.[redacted]	158.[redacted]	17.[redacted]	157.[redacted]	19.[redacted]	25.[redacted]

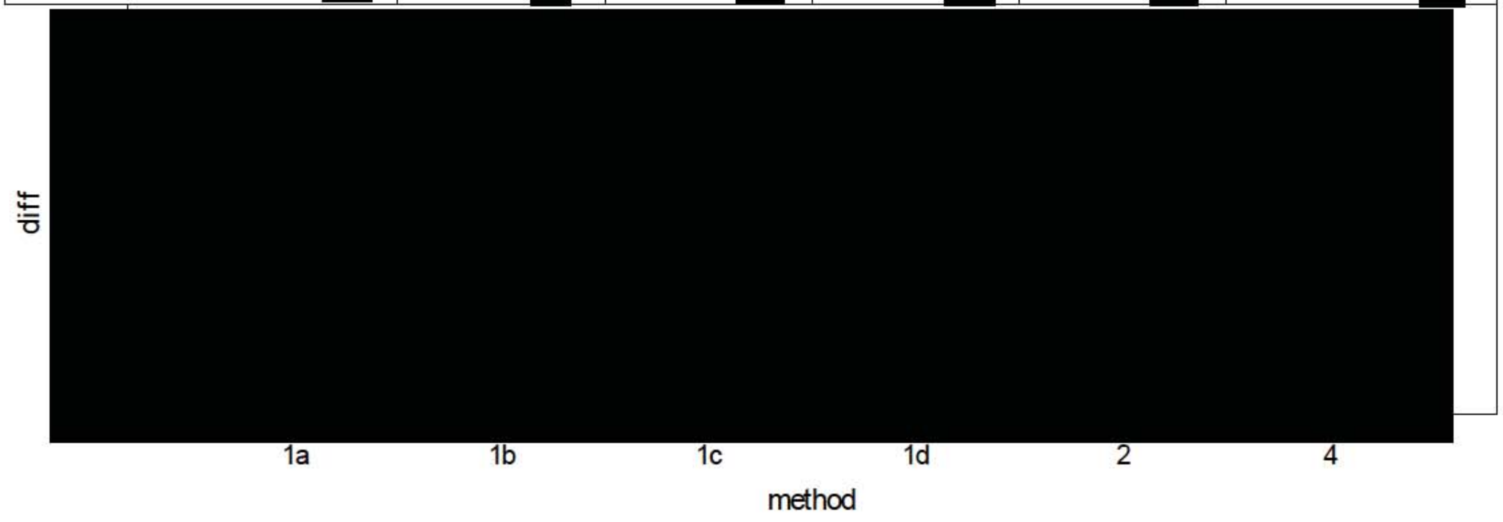

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=203

Overall Statistics					
Min	[redacted]	Mean	-1.[redacted]	Max	[redacted]
Pooled Std Dev	31.[redacted]				

Statistics by GQtype						
N	600	600	600	600	600	600
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-2.[redacted]	-3.[redacted]	-2.[redacted]	-3.[redacted]	-0.[redacted]	0.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	550	550	700	550	300	300
Std Dev	32.[redacted]	32.[redacted]	39.[redacted]	31.[redacted]	22.[redacted]	27.[redacted]

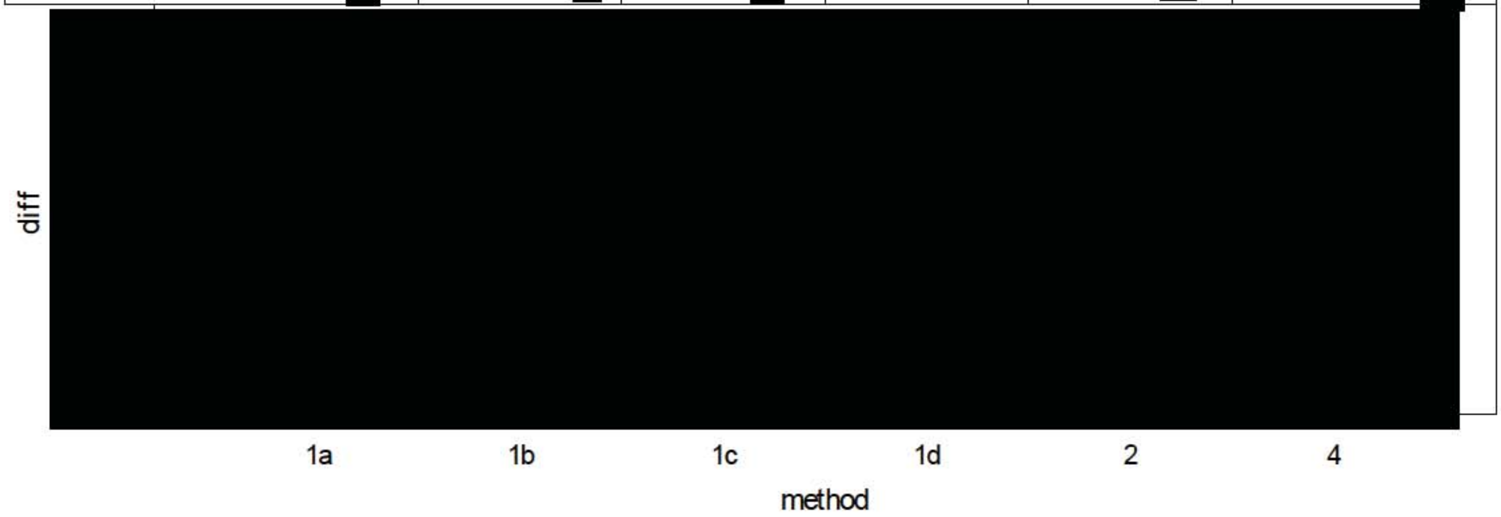

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=301

Overall Statistics					
Min	[redacted]	Mean	-3.[redacted]	Max	[redacted]
Pooled Std Dev	386.[redacted]				

Statistics by GQtype						
N	14500	14500	14500	14500	14500	14500
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-11.[redacted]	0.[redacted]	-11.[redacted]	-0.[redacted]	0.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	1100	75000	1400	75000	1200	1400
Std Dev	25.[redacted]	667.[redacted]	33.[redacted]	668.[redacted]	28.[redacted]	51.[redacted]


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

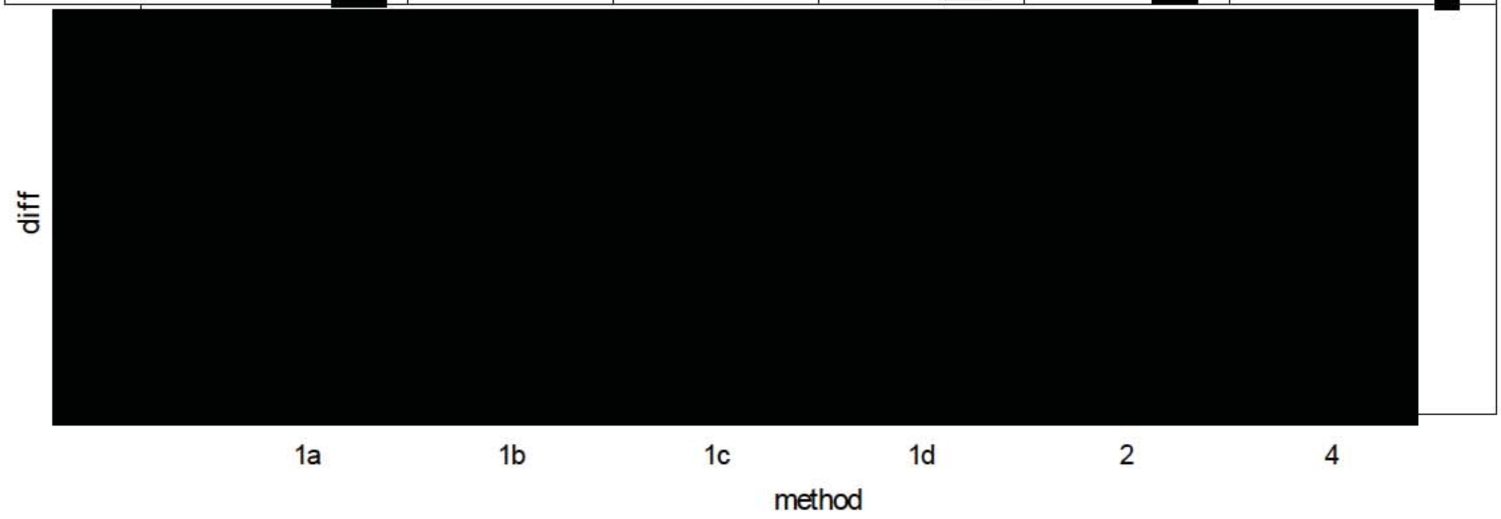
GQTYPCUR=401

Overall Statistics

Min	[redacted]	Mean	1.[redacted]	Max	[redacted]
Pooled Std Dev	93.[redacted]				

Statistics by GQtype

N	250	250	250	250	250	250
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-3.[redacted]	-3.[redacted]	10.[redacted]	-0.[redacted]	-0.[redacted]	6.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	1400	1400	1600	1400	700	1600
Std Dev	77.[redacted]	79.[redacted]	107.[redacted]	80.[redacted]	56.[redacted]	135.[redacted]



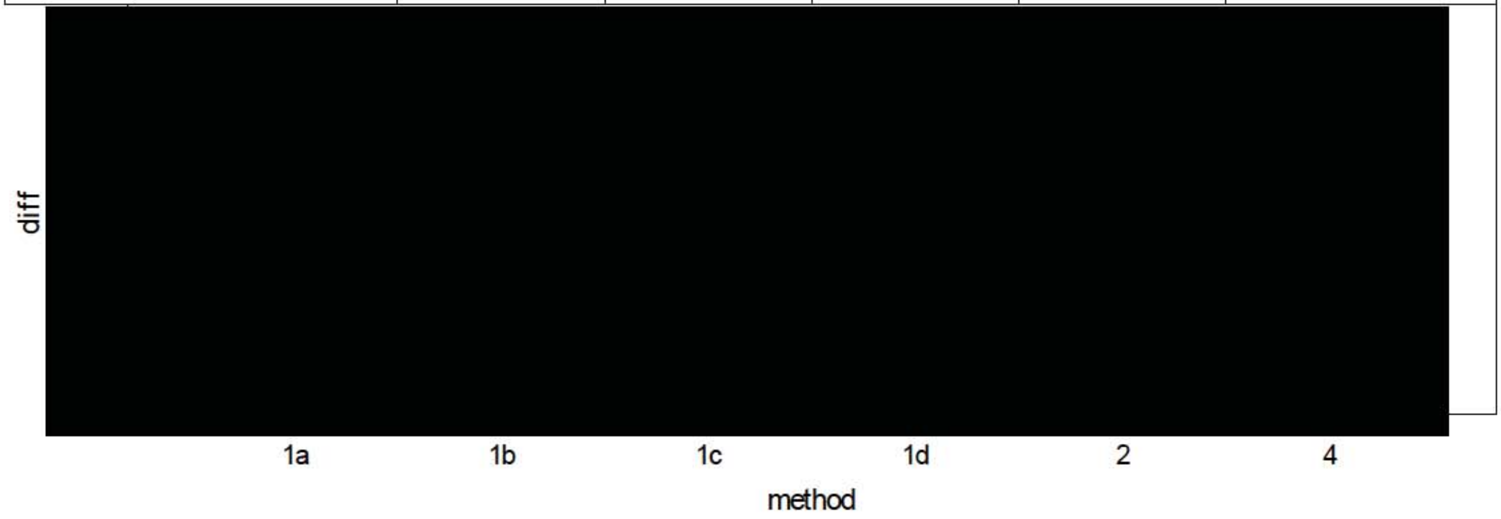
[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=402

Overall Statistics					
Min	[redacted]	Mean	-4.[redacted]	Max	[redacted]
Pooled Std Dev	52				

Statistics by GQtype						
N	70	70	70	70	70	70
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	0.[redacted]	-3	-10	-4	-0.[redacted]	-6
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	150	200	650	400	300	600
Std Dev	20	30	70	40	40	80

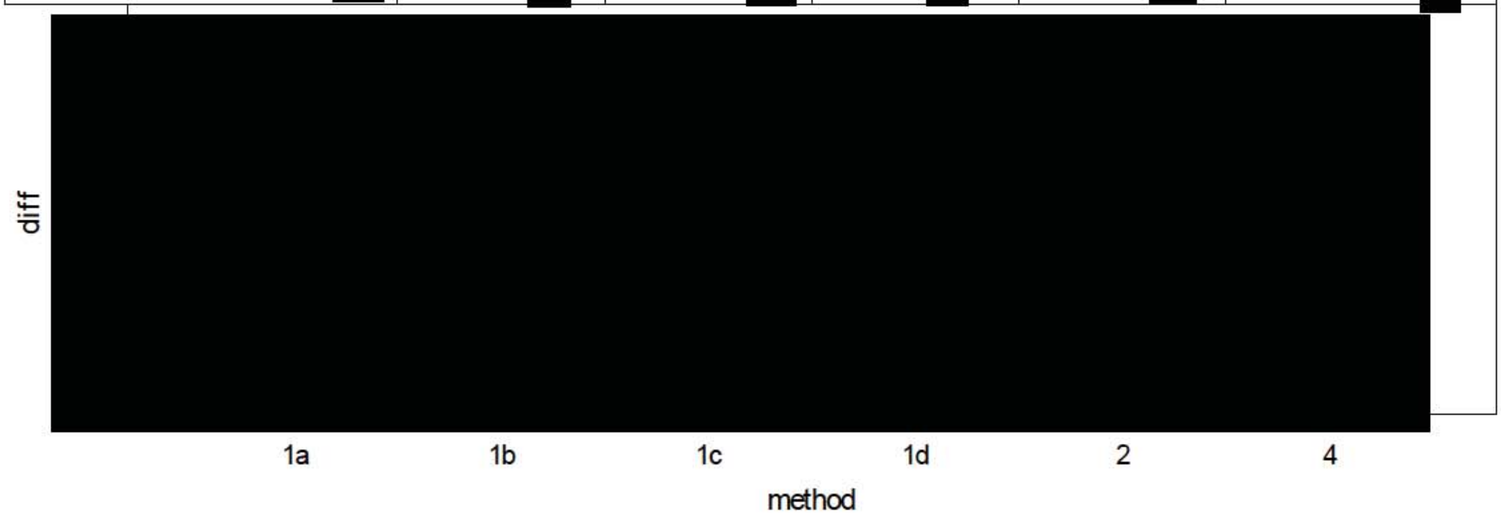

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=403

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	15.[redacted]				

Statistics by GQtype						
N	200	200	200	200	200	200
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	150	100	250	100	150	300
Std Dev	10.[redacted]	10.[redacted]	19.[redacted]	9.[redacted]	10.[redacted]	26.[redacted]


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

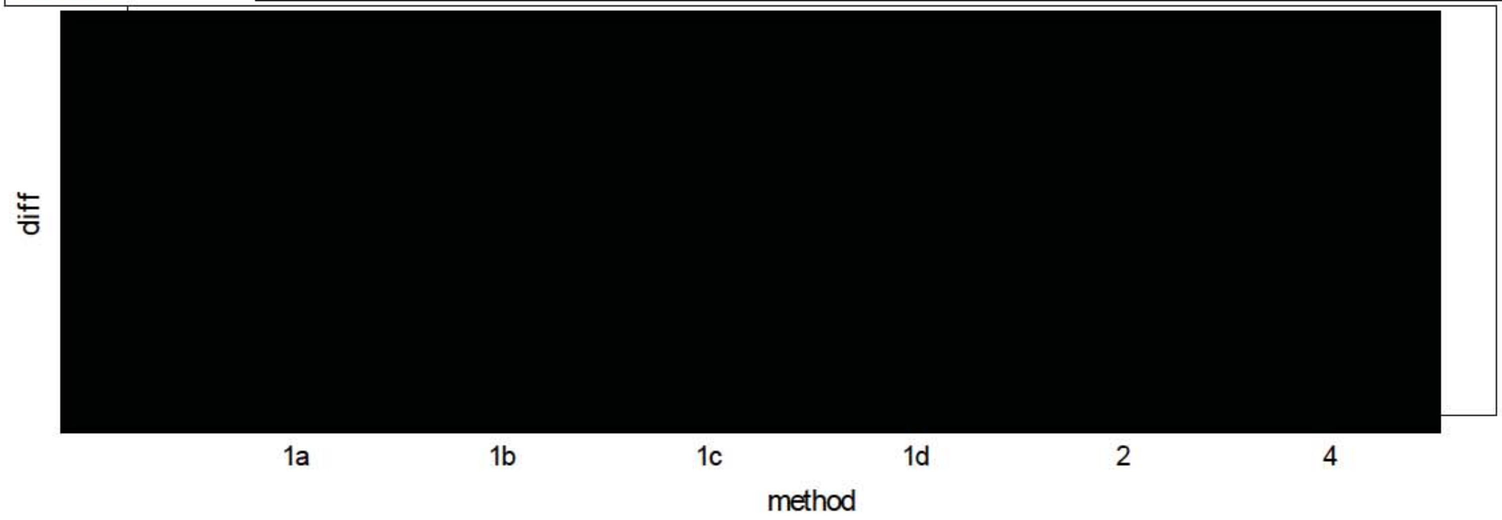
GQTYPCUR=404

Overall Statistics

Min	[redacted]	Mean	-14	Max	[redacted]
Pooled Std Dev	36				

Statistics by GQtype

N	<15	<15	<15	<15	<15	<15
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Std Dev	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

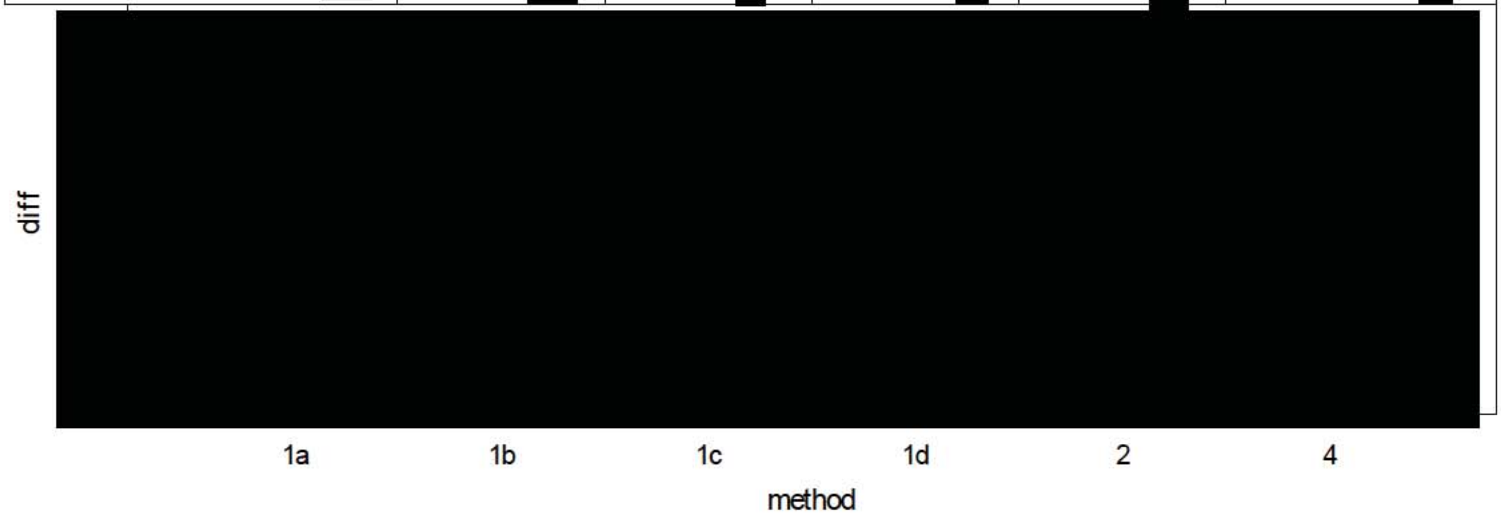

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=405

Overall Statistics					
Min	[redacted]	Mean	-2.[redacted]	Max	[redacted]
Pooled Std Dev	34.[redacted]				

Statistics by GQtype						
N	250	250	250	250	250	250
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-3.[redacted]	-5.[redacted]	0.[redacted]	-5.[redacted]	-0.[redacted]	0.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	250	250	500	250	250	550
Std Dev	27.[redacted]	30.[redacted]	38.[redacted]	29.[redacted]	24.[redacted]	48.[redacted]

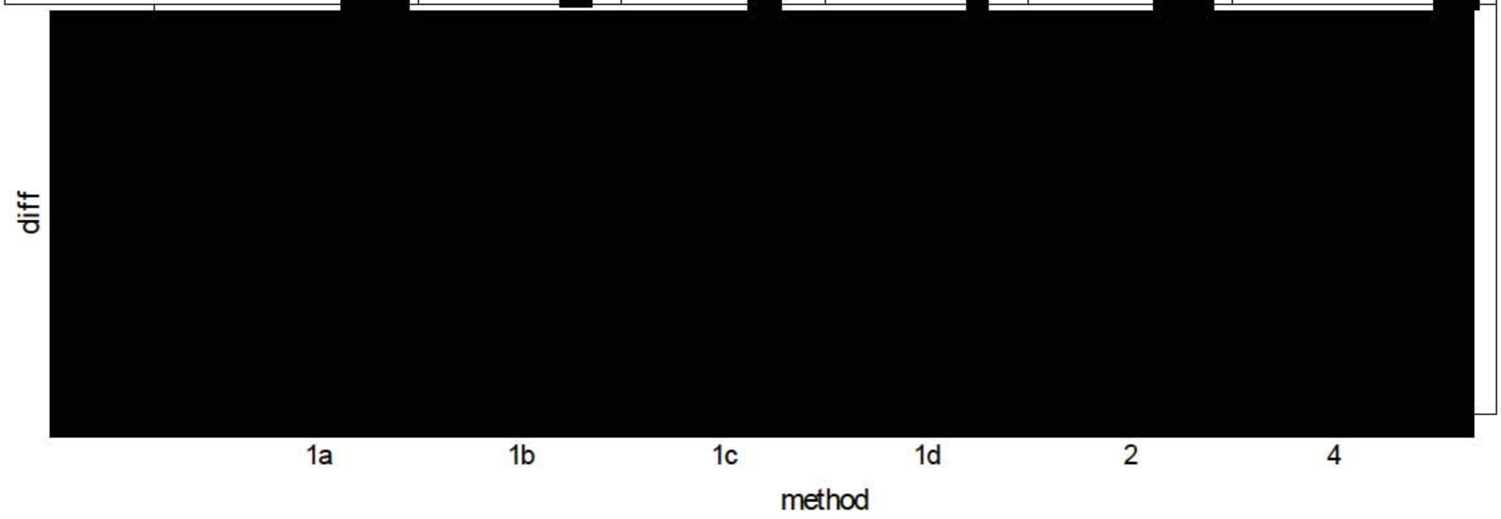

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=501

Overall Statistics					
Min	[redacted]	Mean	-2.[redacted]	Max	[redacted]
Pooled Std Dev	236.[redacted]				

Statistics by GQtype						
N	13000	13000	13000	13000	13000	13000
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean ☺	-0.[redacted]	-6.[redacted]	1.[redacted]	-7.[redacted]	-0.[redacted]	-1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	1200	41000	3500	42500	1600	3000
Std Dev	34.[redacted]	387.[redacted]	55.[redacted]	401.[redacted]	40.[redacted]	131.[redacted]

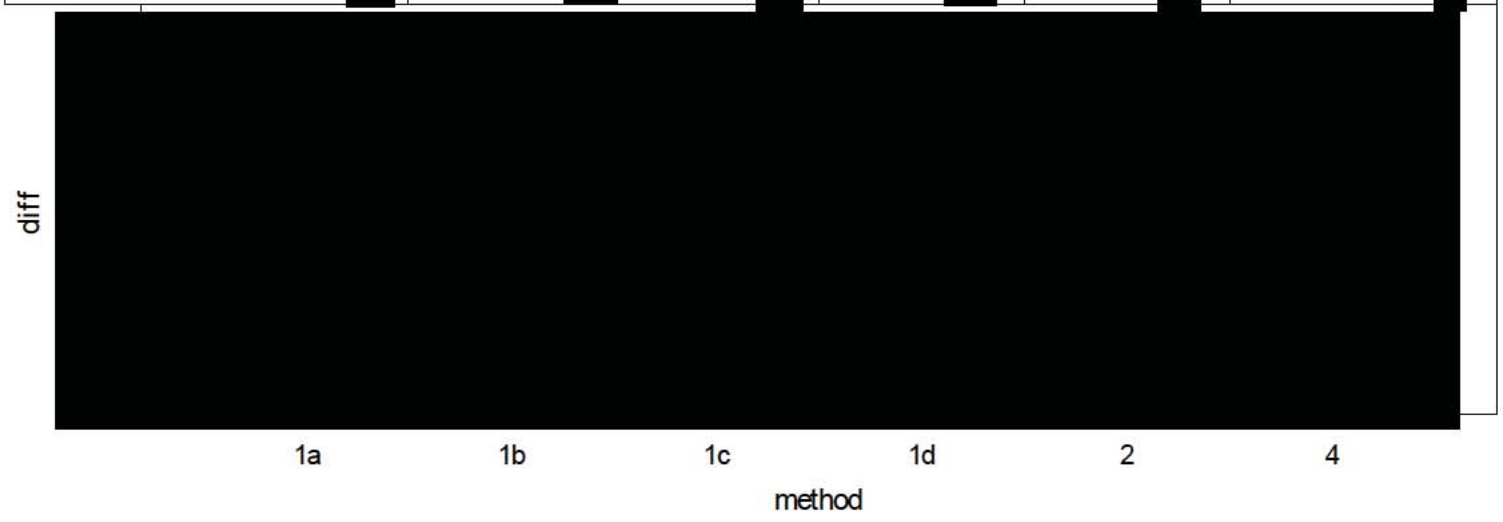

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=601

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	106.[redacted]				

Statistics by GQtype						
N	850	850	850	850	850	850
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean ☺	-1.[redacted]	-4.[redacted]	3.[redacted]	-0[redacted]	0.[redacted]	0.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	1400	1600	2000	1500	1100	1300
Std Dev	93.[redacted]	97.[redacted]	129.[redacted]	94[redacted]	90.[redacted]	127.[redacted]

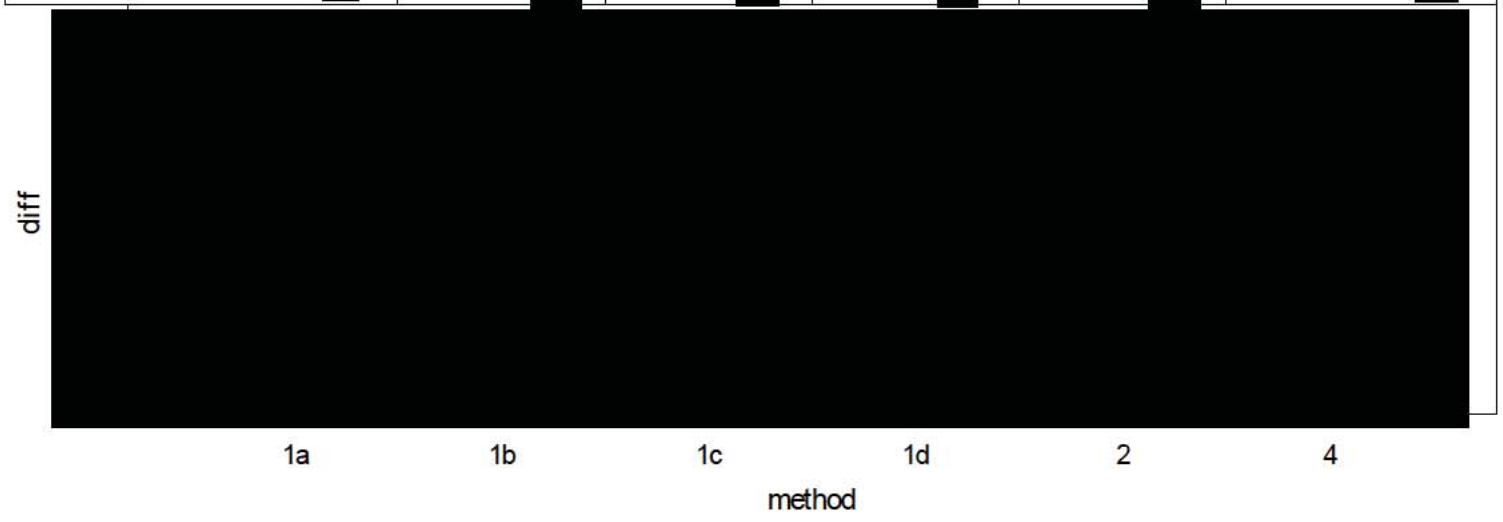

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=701

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	41.[redacted]				

Statistics by GQtype						
N	1600	1600	1600	1600	1600	1600
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-1.[redacted]	-1.[redacted]	-0.[redacted]	-3.[redacted]	-0.[redacted]	4.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	900	800	750	800	600	1200
Std Dev	36.[redacted]	35.[redacted]	42.[redacted]	36[redacted]	34.[redacted]	58[redacted]

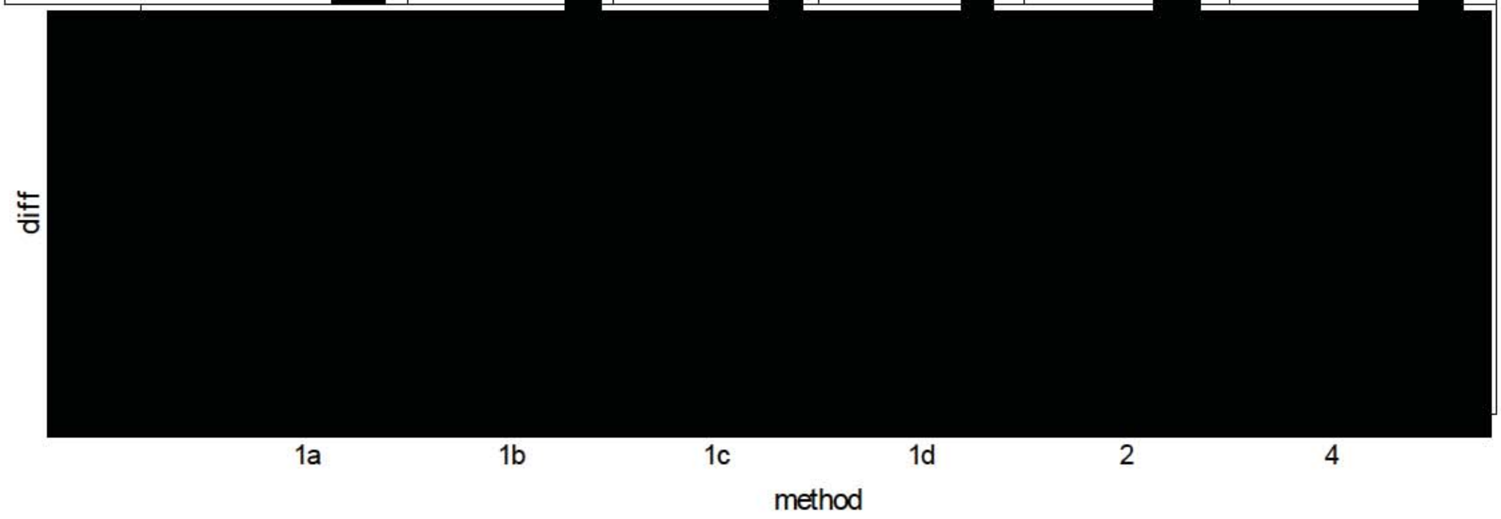

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=702

Overall Statistics					
Min	[redacted]	Mean	-7.[redacted]	Max	[redacted]
Pooled Std Dev	120.[redacted]				

Statistics by GQtype						
N	750	750	750	750	750	750
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2 ☺	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-2.[redacted]	-12.[redacted]	-9.[redacted]	-18.[redacted]	-1.[redacted]	1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	1100	2300	5700	2400	1800	1600
Std Dev	64.[redacted]	112.[redacted]	205.[redacted]	118.[redacted]	78.[redacted]	92.[redacted]

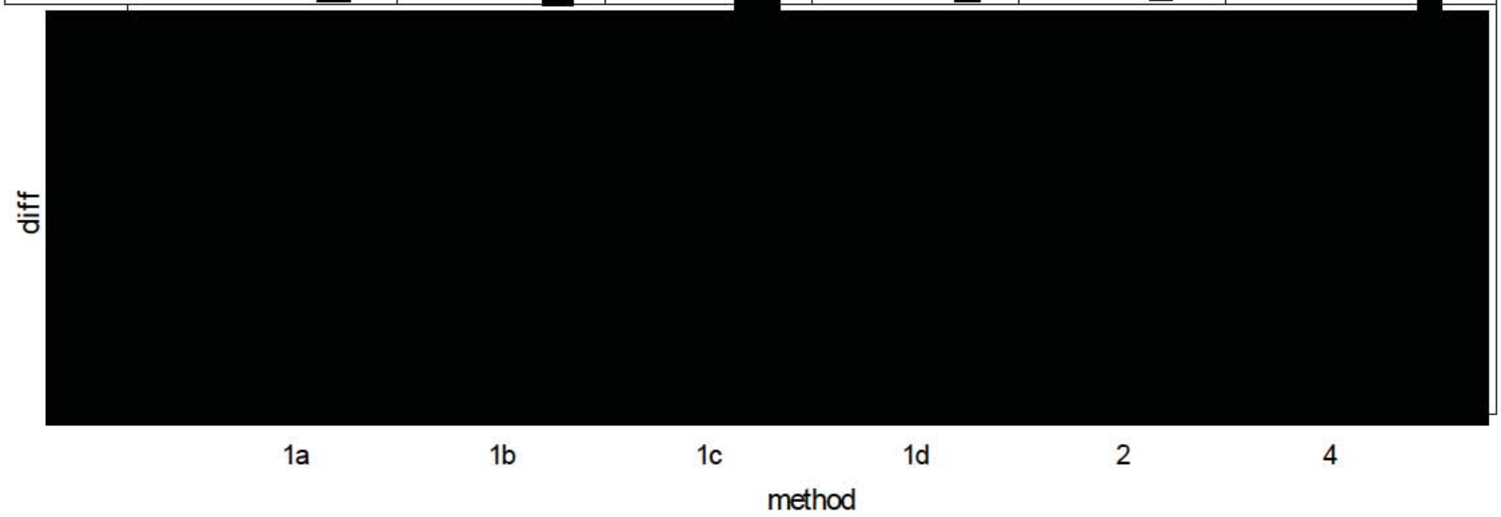

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=704

Overall Statistics					
Min	[redacted]	Mean	-5.[redacted]	Max	[redacted]
Pooled Std Dev	69.[redacted]				

Statistics by GQtype						
N	30	30	30	30	30	30
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	3.[redacted]	-20.[redacted]	-0.[redacted]	-20.[redacted]	-0.[redacted]	6.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	100	750	100	750	90	100
Std Dev	21.[redacted]	116.[redacted]	17.[redacted]	116.[redacted]	17.[redacted]	23.[redacted]


[redacted]

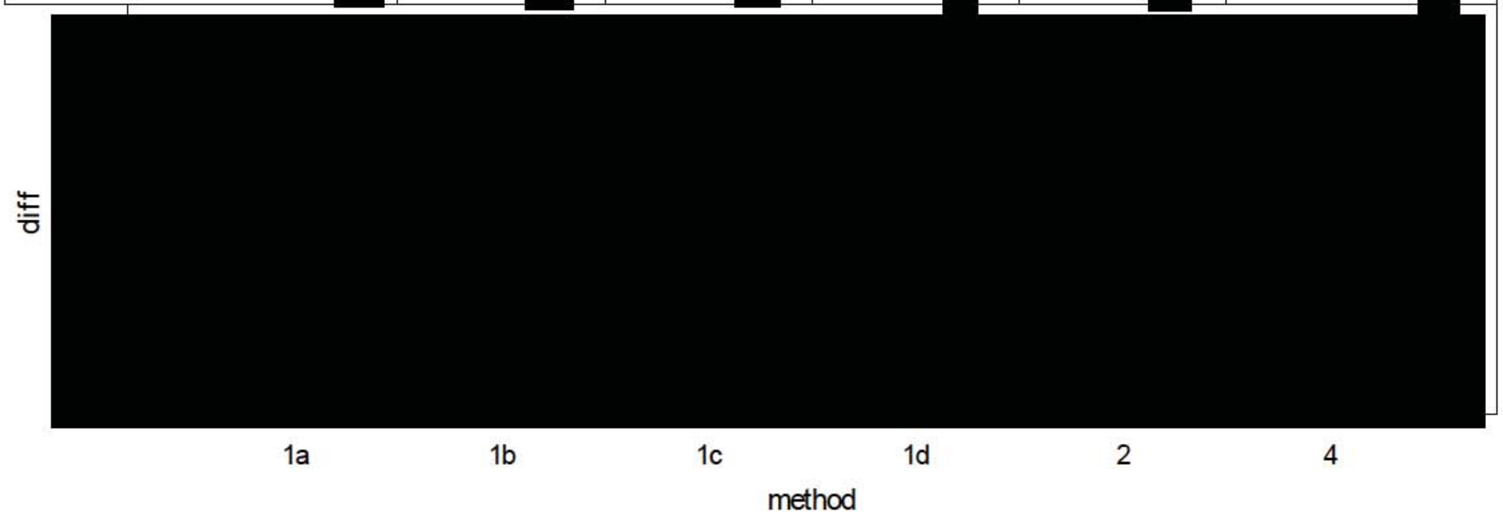
DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=706

Overall Statistics					
Min	[redacted]	Mean	-1.[redacted]	Max	[redacted]
Pooled Std Dev	24.[redacted]				

Statistics by GQtype

N	200	200	200	200	200	200
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-4.[redacted]	-4.[redacted]	-0.[redacted]	-0.[redacted]	1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	150	200	600	150	200	250
Std Dev	18.[redacted]	22.[redacted]	42.[redacted]	16.[redacted]	18.[redacted]	19.[redacted]

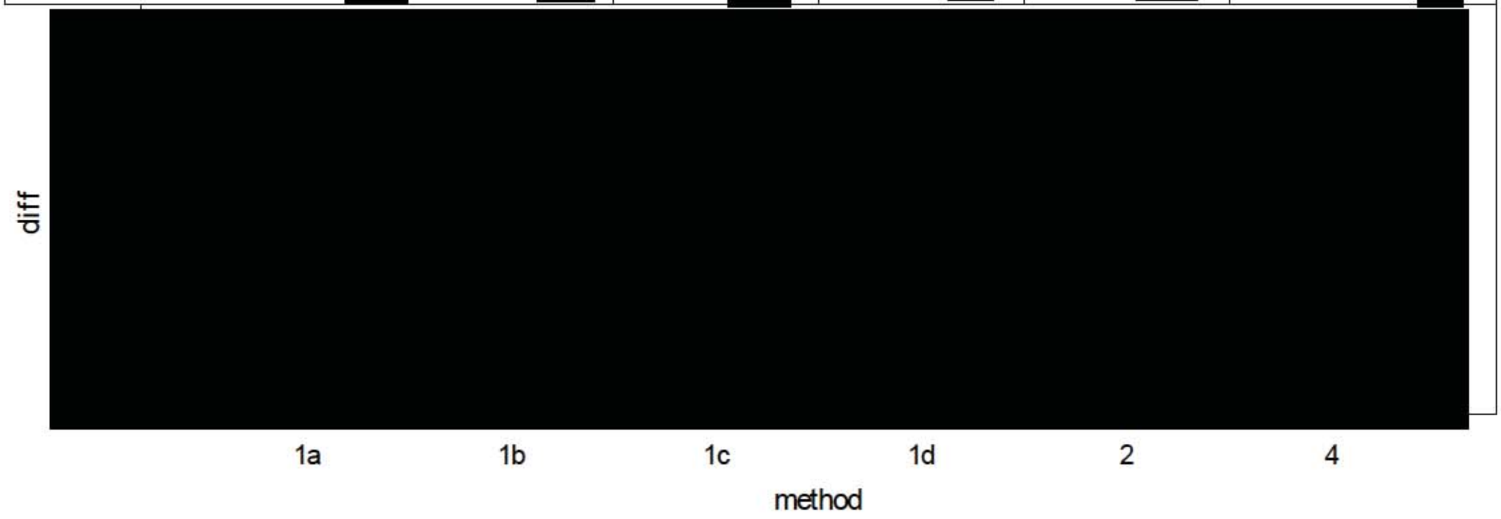

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=801

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	14.[redacted]				

Statistics by GQtype						
N	21000	21000	21000	21000	21000	21000
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	300	2200	650	2100	600	500
Std Dev	5.[redacted]	19.[redacted]	9.[redacted]	19.[redacted]	9.[redacted]	14.[redacted]


[redacted]

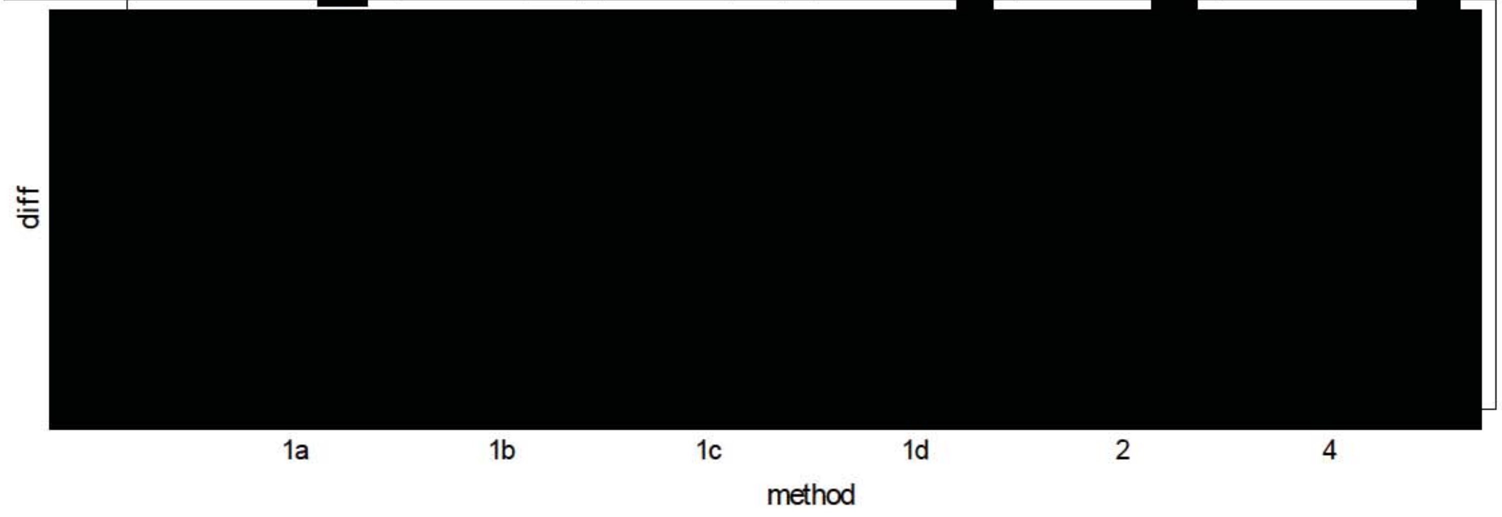
DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=802

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	20.[redacted]				

Statistics by GQtype

N	3600	3600	3600	3600	3600	3600
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	2.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	450	850	650	850	650	550
Std Dev	13.[redacted]	17.[redacted]	20.[redacted]	17.[redacted]	15.[redacted]	31.[redacted]

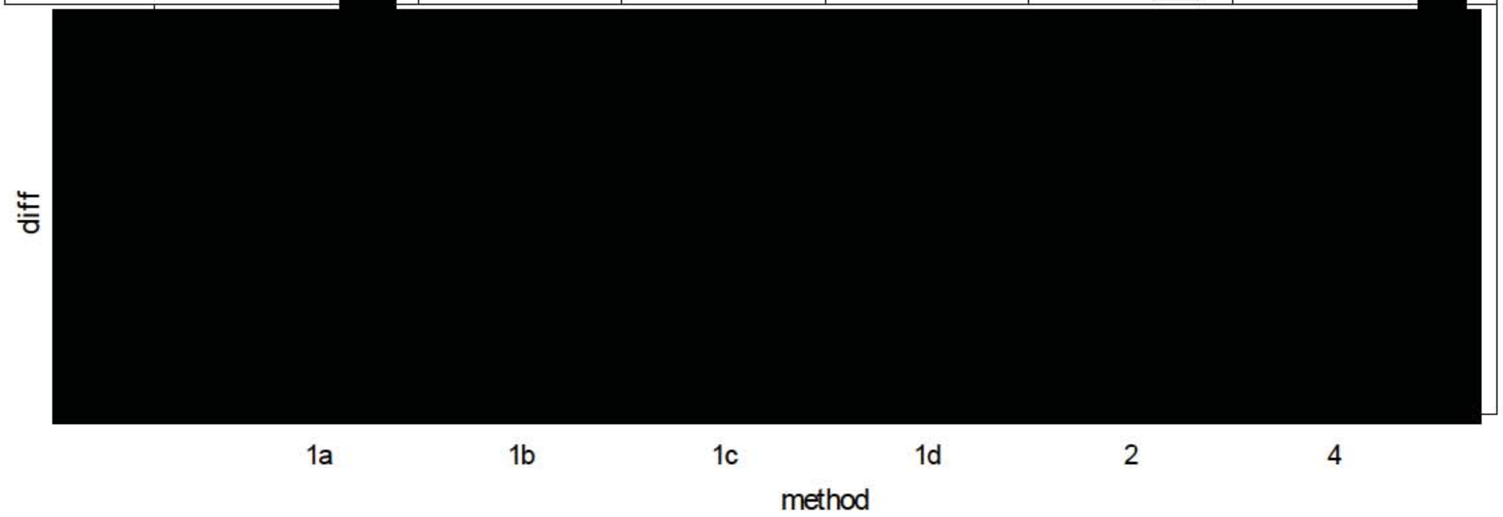

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=901

Overall Statistics					
Min	[redacted]	Mean	-9.[redacted]	Max	[redacted]
Pooled Std Dev	757.[redacted]				

Statistics by GQtype						
N	2300	2300	2300	2300	2300	2300
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-1.[redacted]	-27.[redacted]	1.[redacted]	-30.[redacted]	-0.[redacted]	3.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	500	58500	650	66500	500	600
Std Dev	20.[redacted]	1224	27.[redacted]	1392	20.[redacted]	34.[redacted]



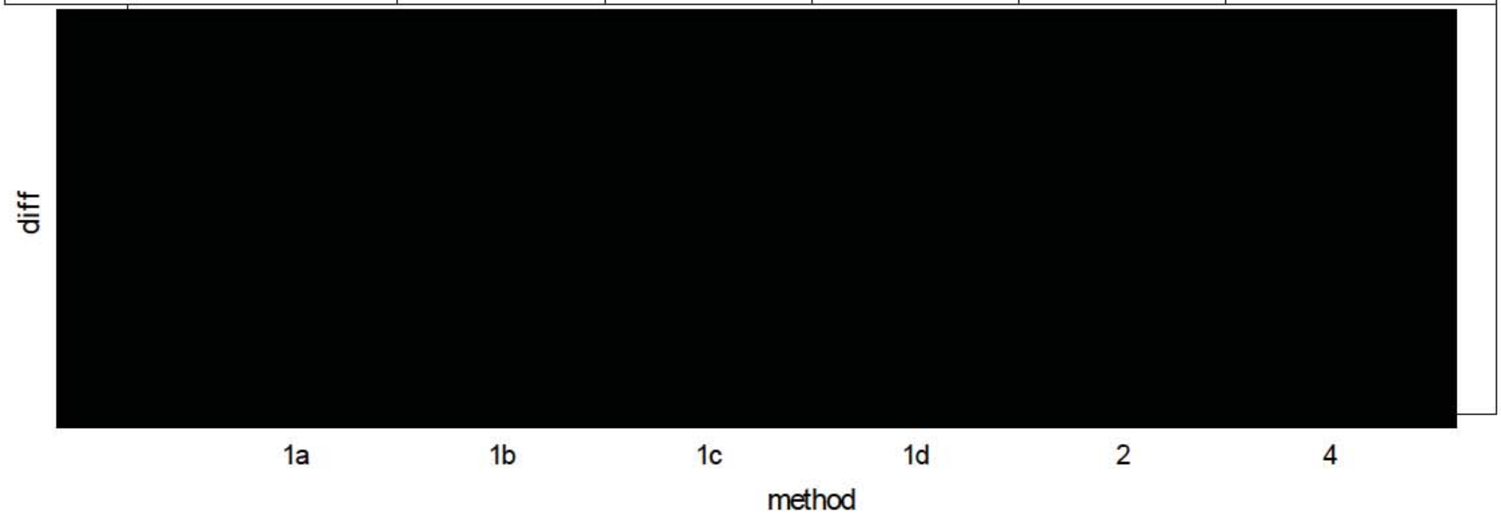
[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=903

Overall Statistics					
Min	[redacted]	Mean	-2.[redacted]	Max	[redacted]
Pooled Std Dev	28.[redacted]				

Statistics by GQtype						
N	<15	<15	<15	<15	<15	<15
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	7	-2	-3	0	0	-18
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	40	<15	<15	<15	<15	150
Std Dev	20	3	5	2	2	67

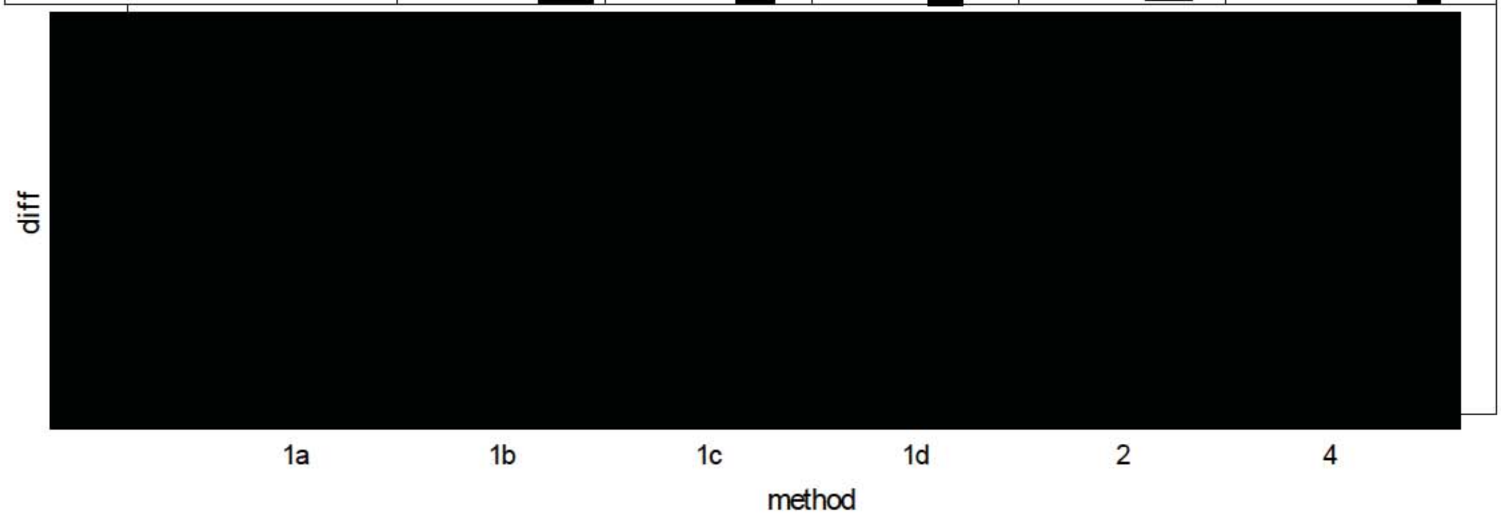

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=999

Overall Statistics					
Min	[redacted]	Mean	-5.[redacted]	Max	[redacted]
Pooled Std Dev	39.[redacted]				

Statistics by GQtype						
N	20	20	20	20	20	20
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-16.[redacted]	-7.[redacted]	-17.[redacted]	-1.[redacted]	-2.[redacted]	12.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	200	90	250	30	200	100
Std Dev	49.[redacted]	20.[redacted]	63.[redacted]	8.[redacted]	38[redacted]	32.[redacted]


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

Variables included in the Group Quarters File for POP Review (gq_mafid_dssd_out_pop) provided by DSSD

Ryan King, Andy Keller, Juli Zamora

December 26, 2020

Variable	Definition	Source	Variable Type and Length
ACOCE	Area Census Office	Universe File	Char $16.
BCUCOUNTYFP	FIPS County Code	Universe File	Char $12.
BCUSTATEFP	FIPS State Code	Universe File	Char $8.
FACTLNAME	The name of the Group Quarters facility. A facility is an umbrella organization that has a group of GQs. For example, a college is the facility and its dorms are the GQs.	Universe File	Char $400.
GQ_SIZE_EXP_PERS_CNT	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact (GQAC)	Num 8 6.
GQ_SIZE_MAX_PERS_CNT	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact (GQAC)	Num 8 6.
GQCONTACT	The name of the Group Quarters primary contact person.	Universe File	Char $140.
GQCURRMAXPOP	Maximum number of people at the Group Quarters.	Universe File, Master Address File	Num 8 6.
GQCURRSIZE	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Universe File, Master Address File	Num 8 6.
GQNAME	The name of the Group Quarters, not the facility name.	Universe File	Char $400.
GQTYPCUR	Current Group Quarters Type code.	Universe File	Char $12.
MAFID	Master Address File ID - Permanent MAFUNIT ID	Universe File	Num 8 10.
GQ_INITIAL_STATUS	Initial Status of GQ before Imputation 1 – Occupied 2 – Vacant 3 – Delete	DSSD	Char $1
GQ_INITIAL_UNRES	Initial Unresolved Status before HB Edit and Imputation 0 - Resolved 1 – Unresolved	DSSD	Num 8

GQ_INITIAL_POP	Number of people with good person flag, GP = 1	DSSD	Num 8
IMPUTE_NEEDED	N = No imputation needed	GEO	Char $1.
GP	Number of people with good person flag, GP = 1 after HB Edit	DSSD HB Edit	Num 8
UNRES	Flag to indicate case is unresolved either due to zero pop or implausible pop	DSSD HB Edit	Num 8

FLAGA	Flag for editing ratio (GP/ GQ_SIZE_EXP_PERS_CNT)
	M = ratio is missing
	R = review
	S = suppress from imputation base
	I = impute
		DSSD HB Edit	Char $1.

FLAGB	Flag for editing ratio (GP/ GQ_SIZE_MAX_PERS_CNT)
	M = ratio is missing
	R = review
	S = suppress from imputation base
	I = impute
		DSSD HB Edit	Char $1.

FLAGC	Flag for editing ratio (GP/GQCURRMAXPOP)
	M = ratio is missing
	R = review
	S = suppress from imputation base
	I = impute
		DSSD HB Edit	Char $1.

FLAGD	Flag for editing ratio (GP/GQCURRSIZE)
	M = ratio is missing
	R = review
	S = suppress from imputation base
	I = impute
		DSSD HB Edit	Char $1.

IMP_GP	Final Imputed Count	DSSD GQCI	Num 8
IMP_FLAG	Path Flag for Final Imputed Count.	DSSD GQCI	Num 8
CALL_STATUS	Call Status from GEO calling operation	GEO	
GEO_POP_COUNT	Population Count from GEO calling operation	GEO	Num 8

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A telephone operation is in progress to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that open on Census Day, but vacant during the GQ Enumeration visit (which started in July 2020) require imputation.

In addition, we will impute a pop size for GQs that have a reported Census Day population count that is much smaller than expected. Our initial proposal is to impute when the Census Day population count is 25% of the GQAC expected count, but research into determining (and refining) this threshould is ongoing.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Visit but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with much lower than expected population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe.

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Open on Census Day, Vacant During Visit	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Additionally, some of the 185,000 resolved occupied GQs will be treated as unresolved because their census day population is much lower than expected. The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs as well as any GQs with a much lower than expected population count. Our current threshold for a "low" population count is < 25% of the GQAC expected count. Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ status. Of the resolved GQs, 89,0000 had a GQAC expected count and 90,000 did not. The

unresolved GQs include the 43,000 GQs without a reported count as well as 4,500 that had a large discrepancy between the GQAC expected population and the reported pop size.

Table 2: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Status	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	88,500	88,000	3,900	17,000	197,000
Vacant During Visit, Open on Census Day	1,000	550	300	19,500	21,500
Refusal GQ	350	450	300	6,700	7,800
Total	90,000	89,000	4,500	43,000	227,000

Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 10 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Type	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	9,900	3,100	300	2,800	16,000
Juvenile Facilities	2,300	3,600	300	1,800	8,000
Nursing Facilities*	6,000	19,000	450	3,200	28,500
Hospitals	750	1,100	100	800	2,800
College Housing*	12,000	17,000	1,400	5,500	36,000
Military*	2,100	900	100	1,900	5,000
Shelters	21,000	3,200	550	8,200	33,000
Group Homes	29,000	32,500	850	9,100	72,000
Other	7,100	8,600	500	9,700	26,000
Total	90,000	89,000	4,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

An alternate definition for a low census day population count would be to use 10% of the GQAC Max Number of People. Table 4 shows counts of the resolved and unresolved cases using this alternate threshold by GQ status. Table 5 shows the same information by GQ type. We will examine using the intersection or union of these conditions as well as setting thresholds at different levels to determine which reported counts require imputation.

Table 4: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Status	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	67,000	111,000	2,900	17,000	197,000
Vacant During Visit, Open on Census Day	550	1,000	350	19,500	21,500
Refusal GQ	150	650	300	6,700	7,800
Total	68,000	112,000	3,500	43,000	227,000

Note that 2,400 GQs with the Low Census Day Pop based on the Max Pop also have a Low Census Day Pop using the GQAC Expected Population.

Table 5: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Type	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	5,600	7,200	400	2,800	16,000
Juvenile Facilities	1,600	4,400	150	1,800	8,000
Nursing Facilities*	4,300	20,500	300	3,200	28,500
Hospitals	550	1,300	90	800	2,800
College Housing*	7,800	21,500	1,200	5,500	36,000
Military*	1,500	1500	90	1,900	5,000
Shelters	17,000	7,300	300	8,200	33,000
Group Homes	24,000	38,500	450	9,100	72,000
Other	5,600	10,000	450	9,700	26,000
Total	68,000	112,000	3,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Variables

Table 6 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, and Administrative Records. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Table 6: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR

Additional sources available for college housing GQs include data collected via web-scraping, data from the Integrated Postsecondary Education Data System (IPEDS) and data from the Common Core. These variables are available at the facility level but not for individual MAFIDs.

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) "**capacity utilization**"—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Additional facility-level variables may become available as research continues.

Table 7: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

First, if a pop count is available from the NPC call operation, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will use a combination of the following methods:

1. Ratio Imputation
2. Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 8 shows that 8,600 of the unresolved GQ can be resolved by converting the GQAC expected count to the GQ pop count using the following ratio adjustment.

Table 8: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For each GQ type, we will use the ratio of the reported GQ Census Day count to the GQAC expected count to convert the GQAC expected count of the unresolved GQ to a Census Day imputed count. For each GQ type, we will calculate the ratio of the sum of the GQAC Expected Count to the sum of the reported GQ population for the resolved cases. For the unresolved GQs, we will multiply the GQAC expected count by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. We will not use ratio imputation with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count. Conversion factors for the four variables under consideration are

shown in Table 9. Tables 12-14 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 9: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0.3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1.017	0.6385
College Housing	0.7818	0.5492	0.9444	0.5535
Military	0.7317	0.2290	0.9492	0.2914
Shelters	0.6261	0.5325	0.6180	0.5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0.3597
All GQs	0.7878	0.5057	0.9217	0.5153

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for GQs for colleges and universities (GQTYPCUR=501).

First, we will adjust the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population, GQAC Max Number of People, and Greek Housing variables.

After adjusting the college-level total room capacity to account reference year and for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count$$
$$- \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Once we calculate the college-level residual, we will then allocate the population counts among the GQs in the college without GQAC Expected Count.

Modeling

A third approach would be to impute the GQ pop counts from a Poisson regression model. The dependent variable will be reported GQ pop count with an offset of the max number of people (because that is filled the most). Independent variables will be selected from Table 6. It is important to note that GQ type will either be a fixed-effect covariate in the models or separate models will be fit by GQ type.

Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

Median Imputation

If sufficient auxiliary data is not available, we will impute the pop size with median population within an imputation cell. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and GQ status. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median GQ pop size in the cell.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

We will evaluate the imputation methods using cross validation. First, we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The remaining 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will impute the GQ pop size for all GQs in the validation deck. That is, we will attempt to impute the GQ pop size for every GQ in the 60% sample four times (once for each of the four methods). Note that the second method can only be applied to college housing. Then, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each method. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Some methods may perform better than others for certain types of units. For example, Poisson regression might perform best when the GQAC expected count is available, but not well when it is missing. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Appendix

Table 10: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Table 11: GQAC Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Table 12: GQAC Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Table 13: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Table 14: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

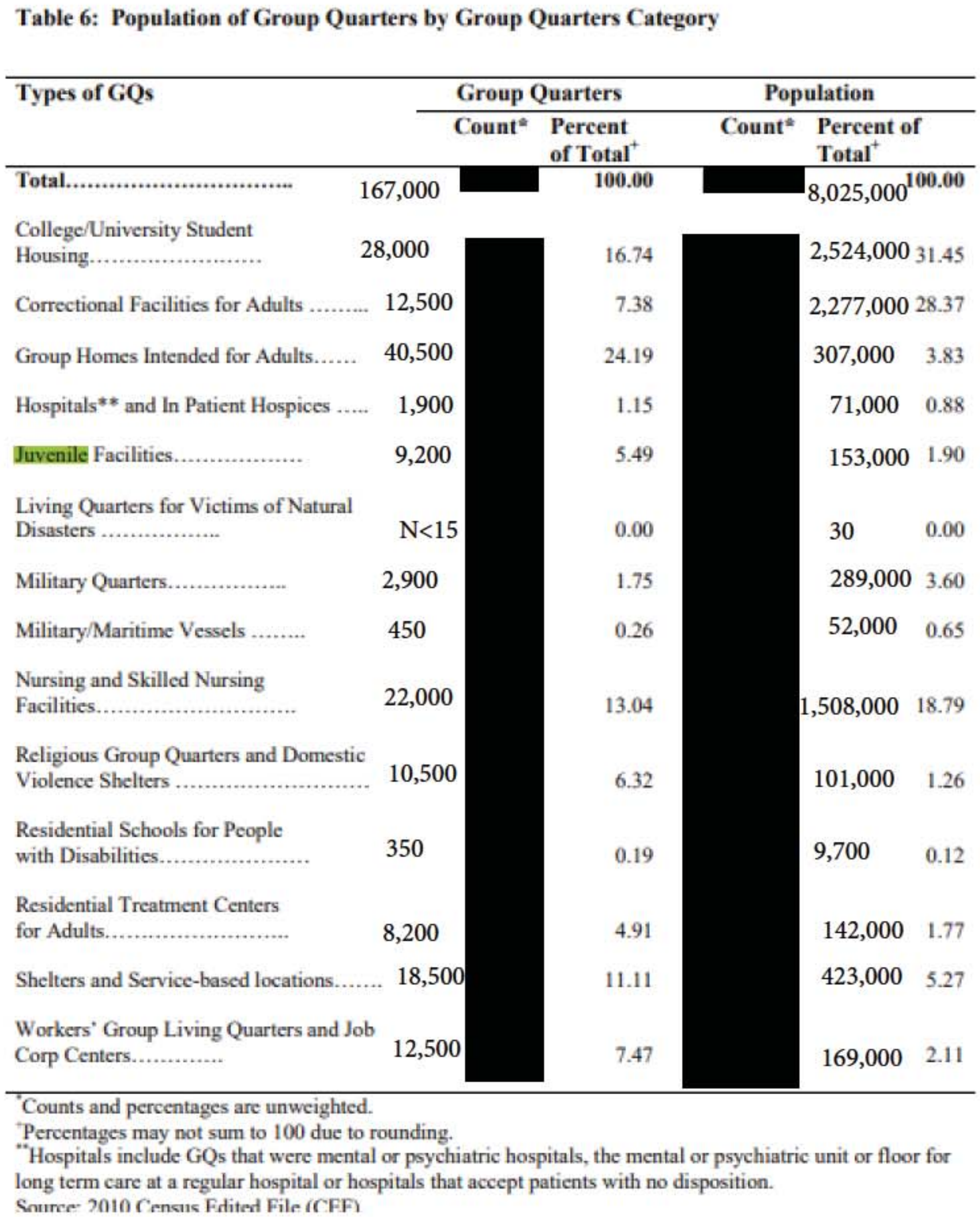


Table 6: Population of Group Quarters by Group Quarters Category

Types of GQs	Group Quarters		Population	
	Count*	Percent of Total+	Count*	Percent of Total+
Total..............................	167,000 [redacted]	**100.00**	[redacted] 8,025,000	**100.00**
College/University Student Housing........................	28,000 [redacted]	16.74	[redacted] 2,524,000	31.45
Correctional Facilities for Adults	12,500 [redacted]	7.38	[redacted] 2,277,000	28.37
Group Homes Intended for Adults......	40,500 [redacted]	24.19	[redacted] 307,000	3.83
Hospitals** and In Patient Hospices	1,900 [redacted]	1.15	[redacted] 71,000	0.88
Juvenile Facilities..................	9,200 [redacted]	5.49	[redacted] 153,000	1.90
Living Quarters for Victims of Natural Disasters	N<15 [redacted]	0.00	[redacted] 30	0.00
Military Quarters.................	2,900 [redacted]	1.75	[redacted] 289,000	3.60
Military/Maritime Vessels	450 [redacted]	0.26	[redacted] 52,000	0.65
Nursing and Skilled Nursing Facilities............................	22,000 [redacted]	13.04	[redacted] 1,508,000	18.79
Religious Group Quarters and Domestic Violence Shelters	10,500 [redacted]	6.32	[redacted] 101,000	1.26
Residential Schools for People with Disabilities.....................	350 [redacted]	0.19	[redacted] 9,700	0.12
Residential Treatment Centers for Adults.........................	8,200 [redacted]	4.91	[redacted] 142,000	1.77
Shelters and Service-based locations.......	18,500 [redacted]	11.11	[redacted] 423,000	5.27
Workers' Group Living Quarters and Job Corp Centers............	12,500 [redacted]	7.47	[redacted] 169,000	2.11

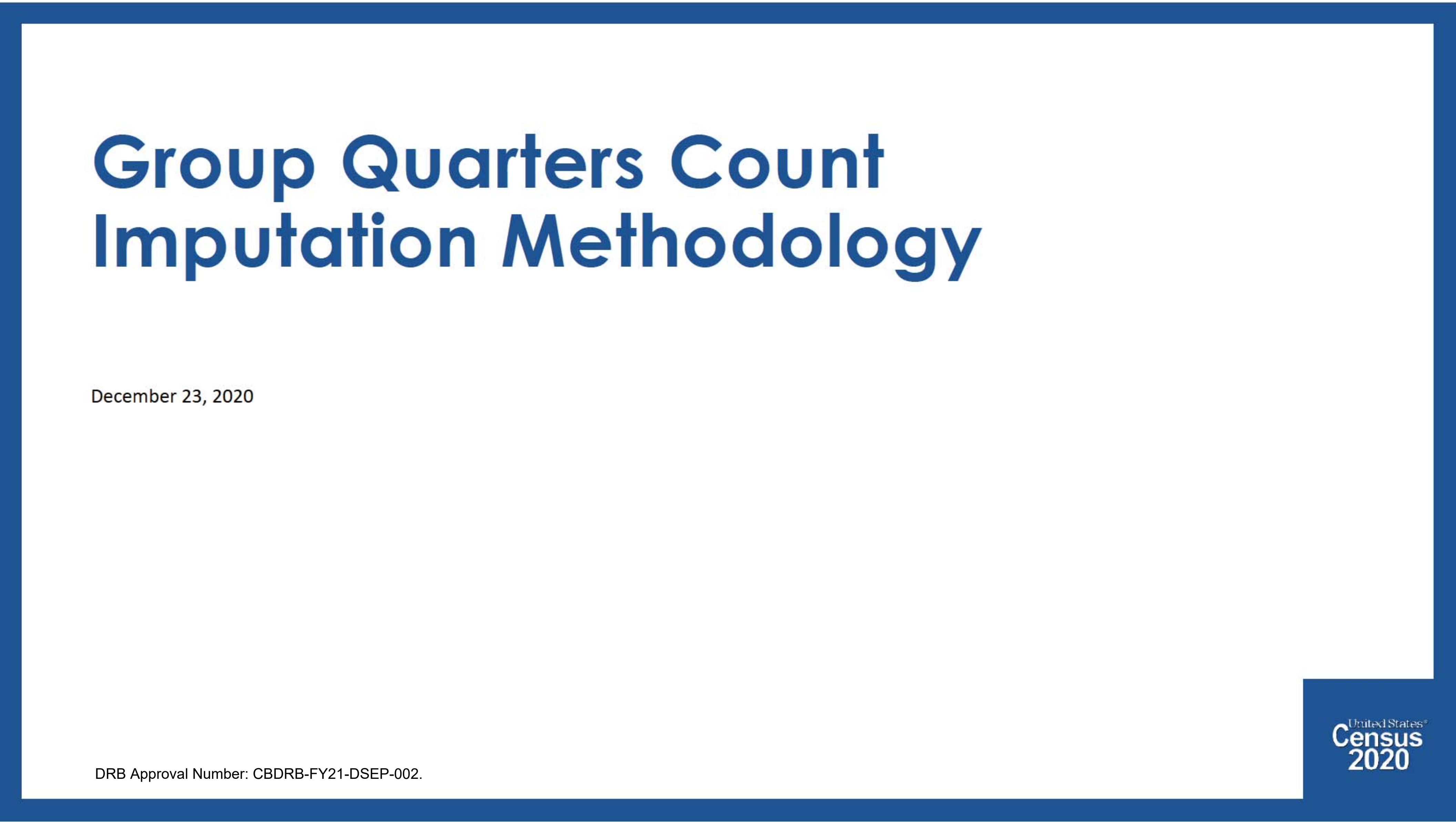
*Counts and percentages are unweighted.

+Percentages may not sum to 100 due to rounding.

**Hospitals include GQs that were mental or psychiatric hospitals, the mental or psychiatric unit or floor for long term care at a regular hospital or hospitals that accept patients with no disposition.

Source: 2010 Census Edited File (CEF)

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



Group Quarters Count Imputation Methodology

December 23, 2020

DRB Approval Number: CBDRB-FY21-DSEP-002.

United States® Census 2020

[redacted]

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002.

United States® Census 2020

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002.

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002.

United States® Census 2020

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002.

[redacted]

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002.

[redacted]

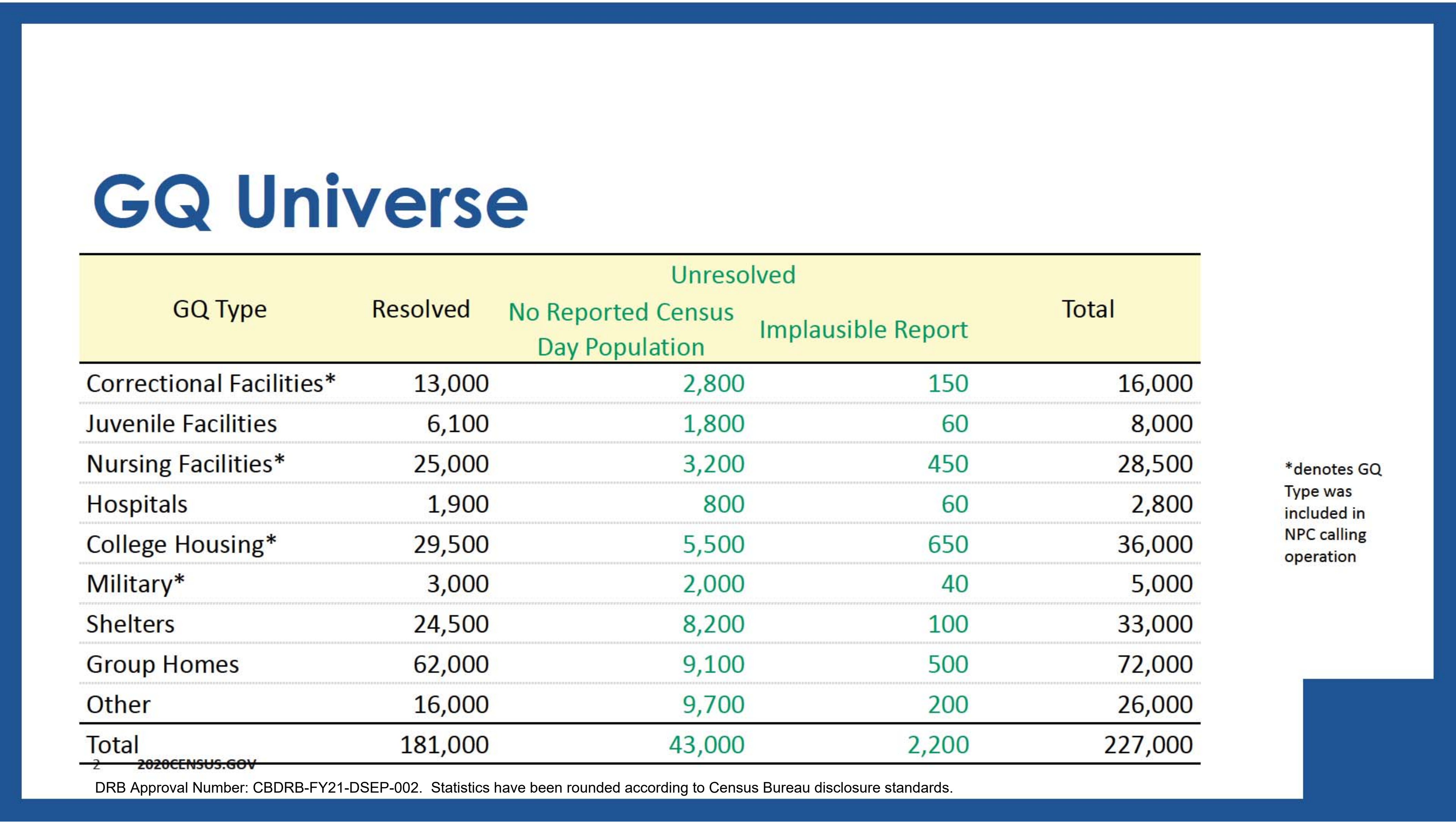
DRB Approval Number: CBDRB-FY21-DSEP-002.

[redacted]

Group Quarters Count Imputation Methodology

December 23, 2020

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

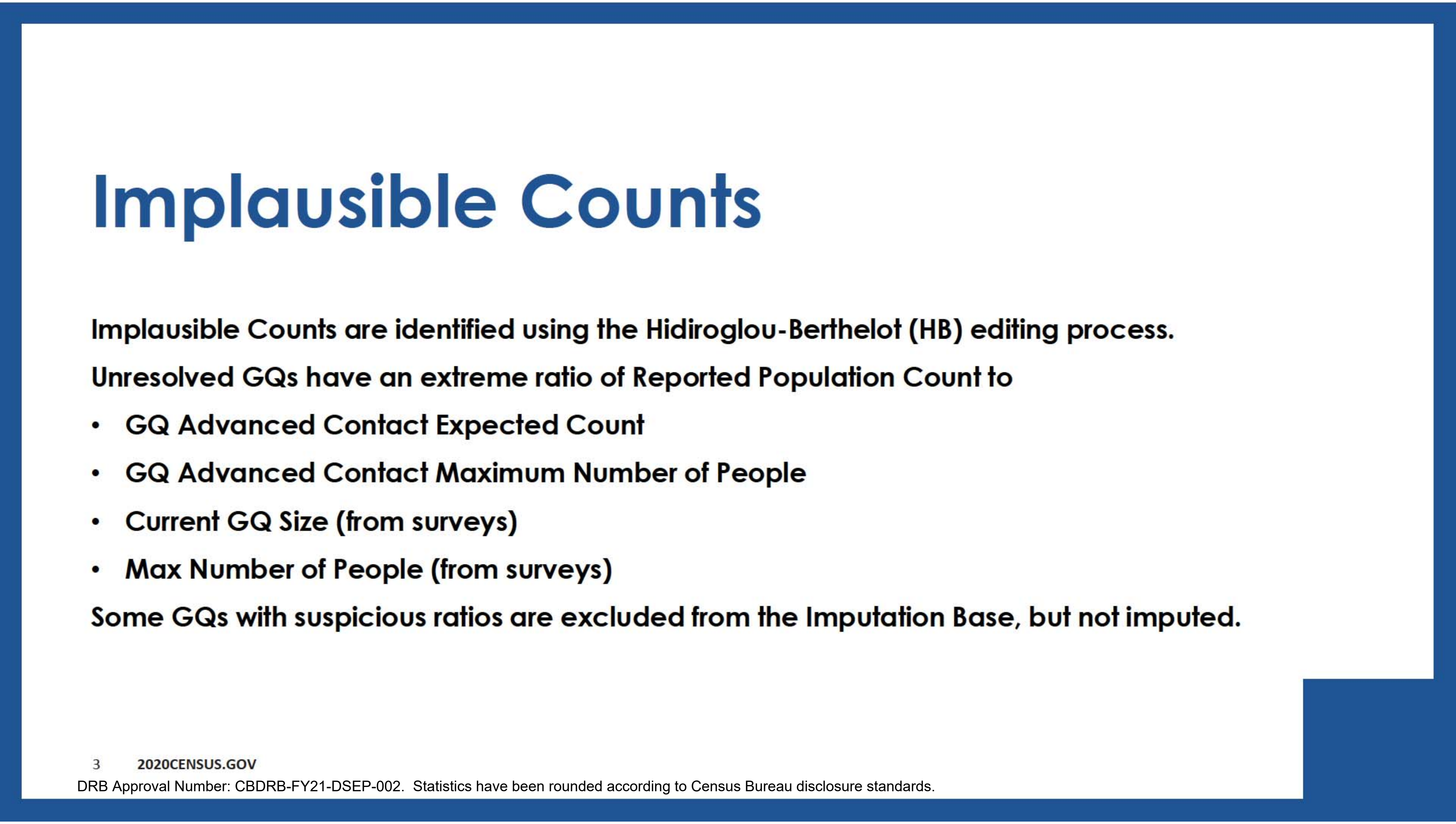


GQ Universe

GQ Type	Resolved	Unresolved		Total
		No Reported Census Day Population	Implausible Report	
Correctional Facilities*	13,000	2,800	150	16,000
Juvenile Facilities	6,100	1,800	60	8,000
Nursing Facilities*	25,000	3,200	450	28,500
Hospitals	1,900	800	60	2,800
College Housing*	29,500	5,500	650	36,000
Military*	3,000	2,000	40	5,000
Shelters	24,500	8,200	100	33,000
Group Homes	62,000	9,100	500	72,000
Other	16,000	9,700	200	26,000
Total	181,000	43,000	2,200	227,000

*denotes GQ Type was included in NPC calling operation

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



Implausible Counts

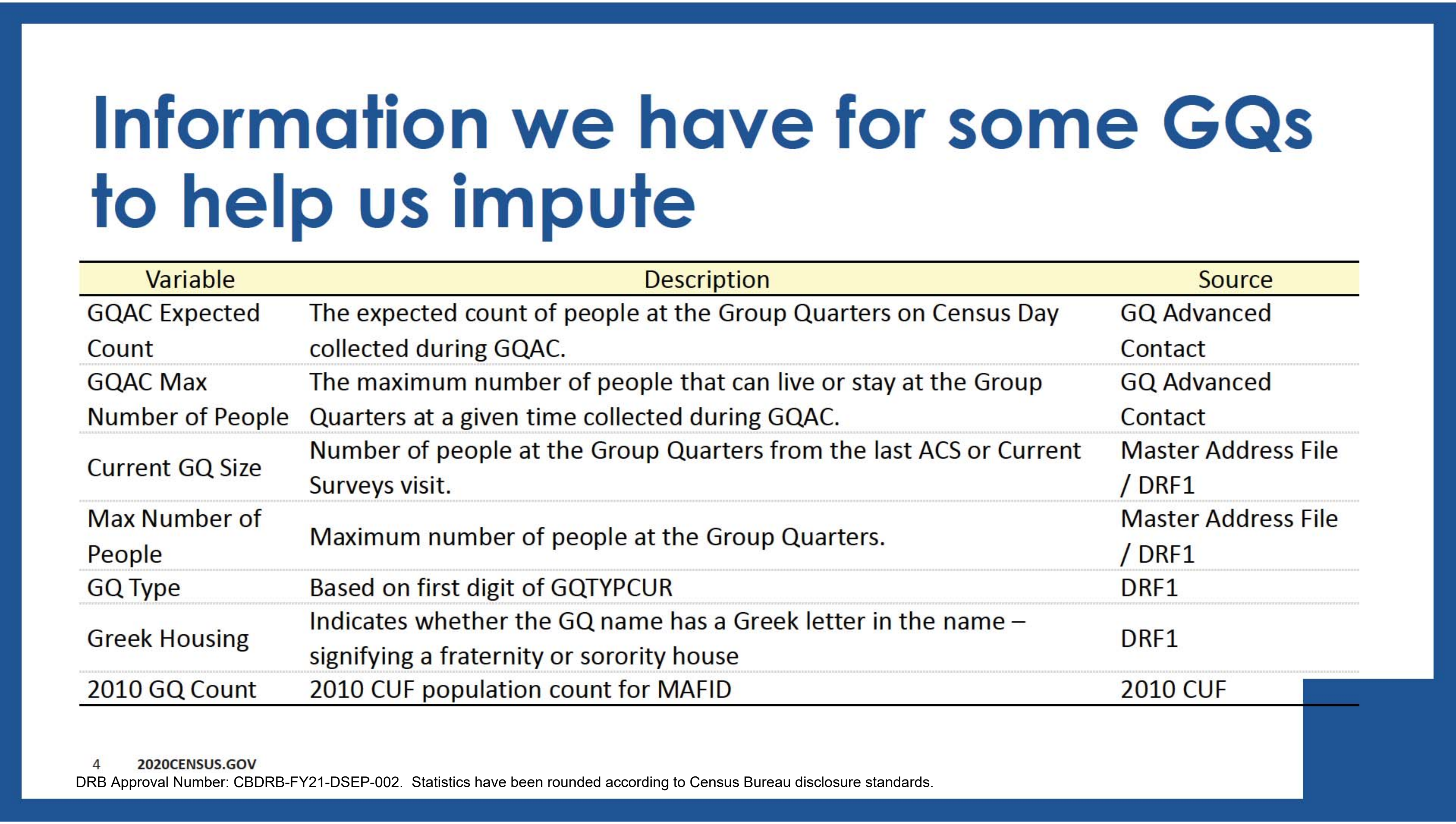
Implausible Counts are identified using the Hidiroglou-Berthelot (HB) editing process.

Unresolved GQs have an extreme ratio of Reported Population Count to

- **GQ Advanced Contact Expected Count**
- **GQ Advanced Contact Maximum Number of People**
- **Current GQ Size (from surveys)**
- **Max Number of People (from surveys)**

Some GQs with suspicious ratios are excluded from the Imputation Base, but not imputed.

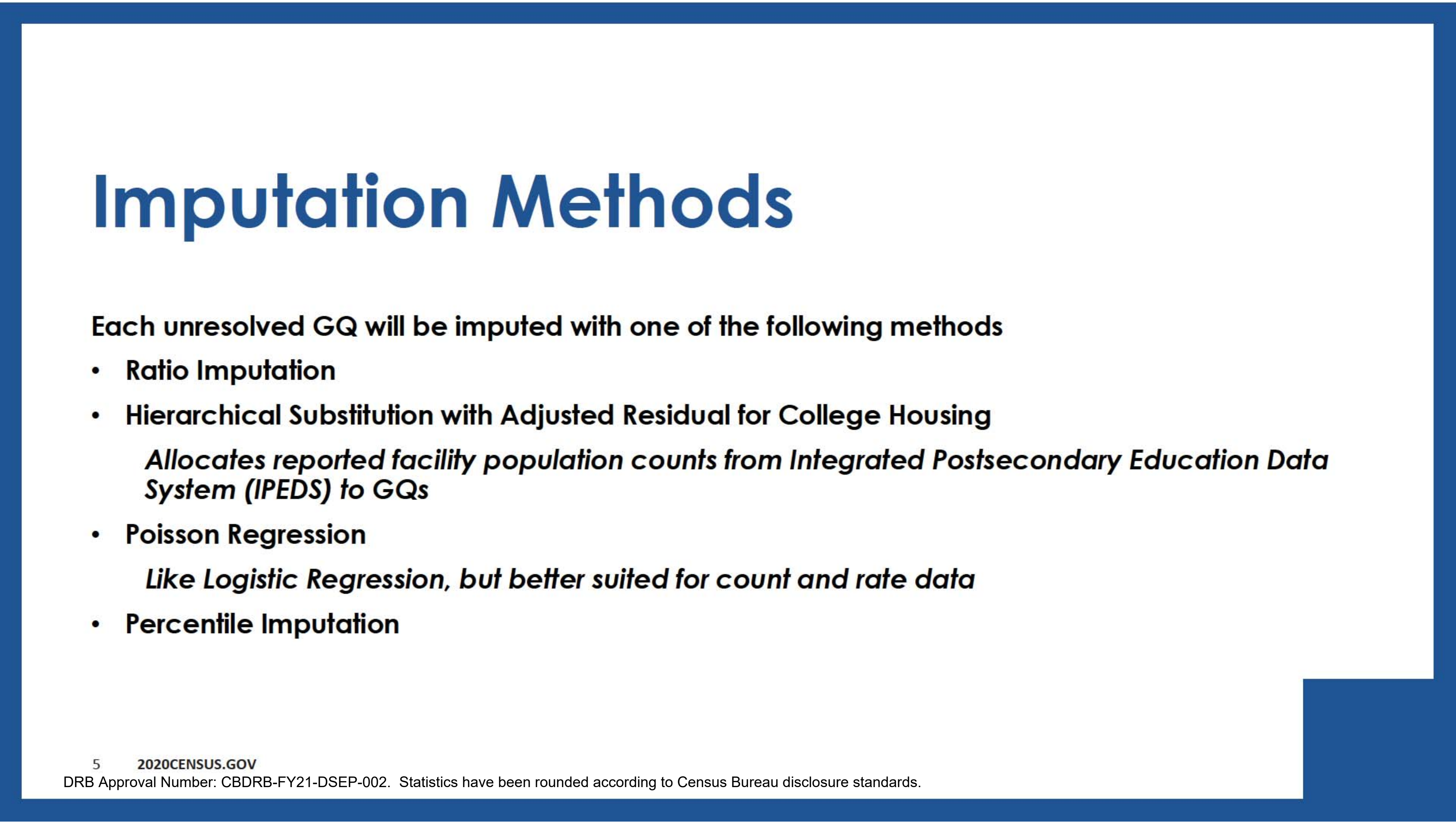
DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



Information we have for some GQs to help us impute

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



Imputation Methods

Each unresolved GQ will be imputed with one of the following methods

- **Ratio Imputation**
- **Hierarchical Substitution with Adjusted Residual for College Housing**

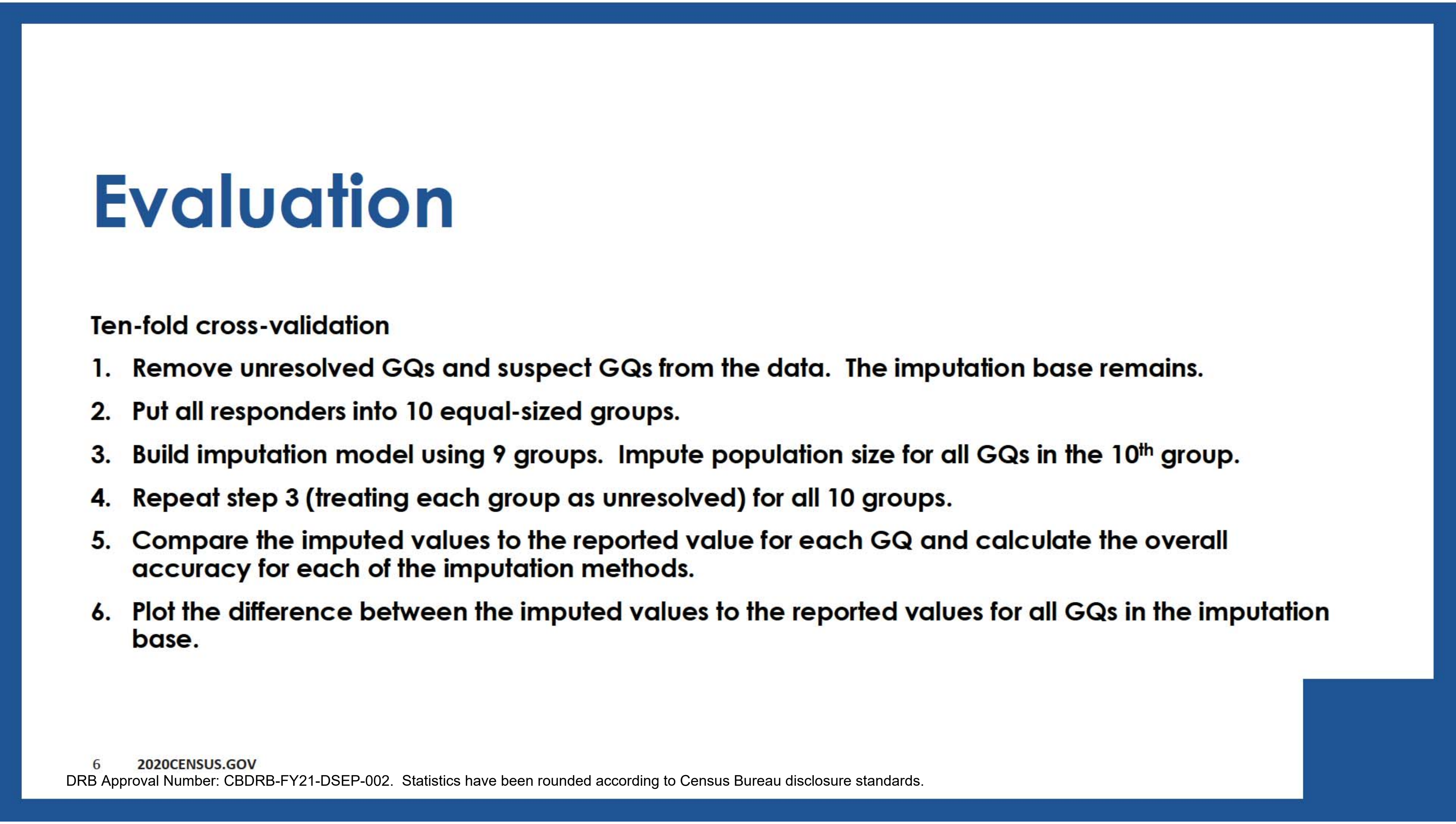
  Allocates reported facility population counts from Integrated Postsecondary Education Data System (IPEDS) to GQs

- **Poisson Regression**

  Like Logistic Regression, but better suited for count and rate data

- **Percentile Imputation**

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



Evaluation

Ten-fold cross-validation

1. Remove unresolved GQs and suspect GQs from the data. The imputation base remains.
2. Put all responders into 10 equal-sized groups.
3. Build imputation model using 9 groups. Impute population size for all GQs in the 10th group.
4. Repeat step 3 (treating each group as unresolved) for all 10 groups.
5. Compare the imputed values to the reported value for each GQ and calculate the overall accuracy for each of the imputation methods.
6. Plot the difference between the imputed values to the reported values for all GQs in the imputation base.

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

[redacted]

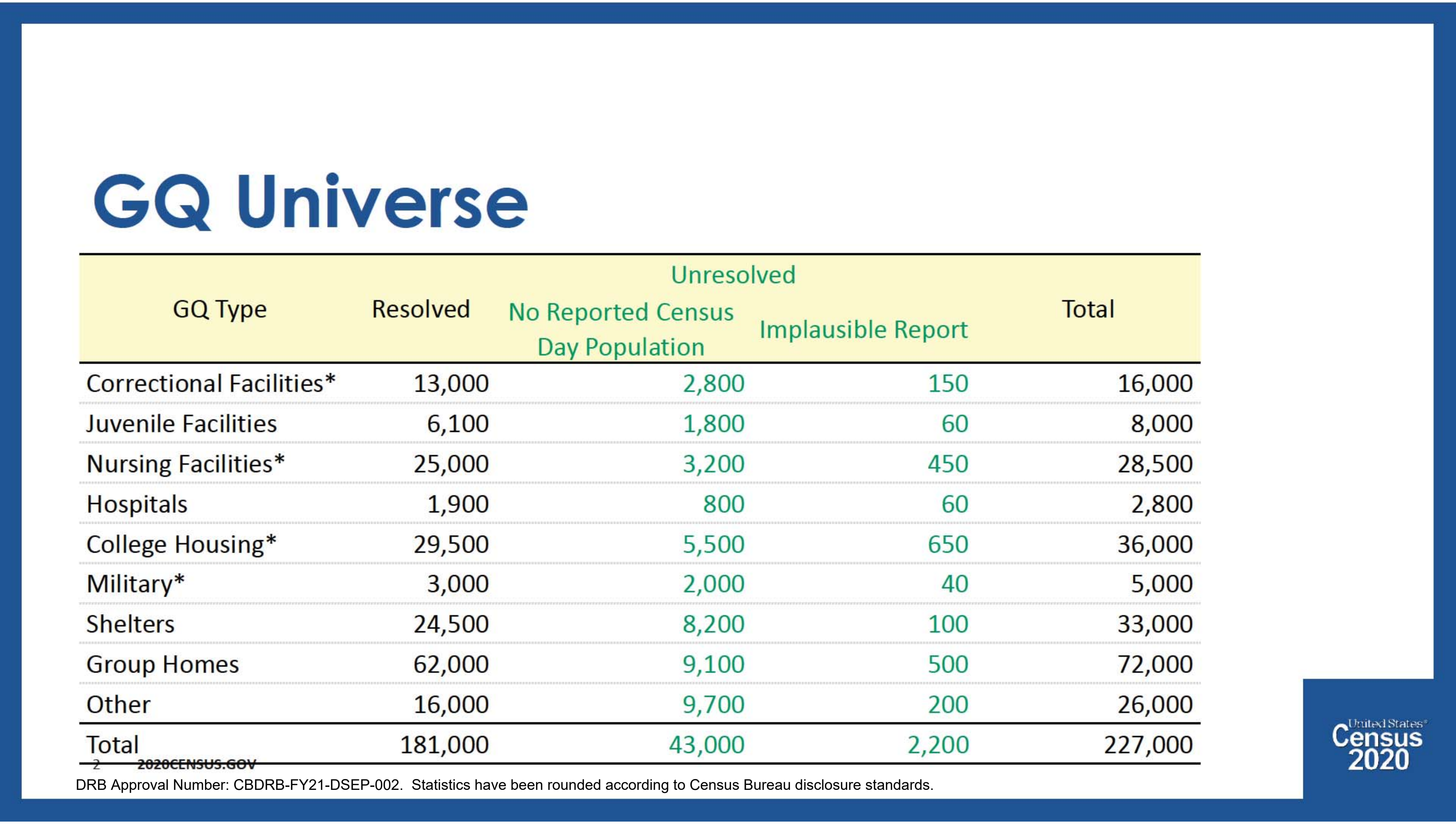
DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

[redacted]

Group Quarters Count Imputation Methodology

December 23, 2020

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



GQ Universe

GQ Type	Resolved	Unresolved		Total
		No Reported Census Day Population	Implausible Report	
Correctional Facilities*	13,000	2,800	150	16,000
Juvenile Facilities	6,100	1,800	60	8,000
Nursing Facilities*	25,000	3,200	450	28,500
Hospitals	1,900	800	60	2,800
College Housing*	29,500	5,500	650	36,000
Military*	3,000	2,000	40	5,000
Shelters	24,500	8,200	100	33,000
Group Homes	62,000	9,100	500	72,000
Other	16,000	9,700	200	26,000
Total	181,000	43,000	2,200	227,000

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Implausible Counts

Implausible Counts are identified using the Hidiroglou-Berthelot (HB) editing process.

Unresolved GQs have an extreme ratio of Reported Population Count to

- GQ Advanced Contact Expected Count
- GQ Advanced Contact Maximum Number of People
- Current GQ Size (from surveys)
- Max Number of People (from surveys)

Some GQs with suspicious ratios are excluded from the Imputation Base, but not imputed.

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Information we have for some GQs to help us impute

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Imputation Methods

Each unresolved GQ will be imputed with one of the following methods

- **Ratio Imputation**
- **Hierarchical Substitution with Adjusted Residual for College Housing**

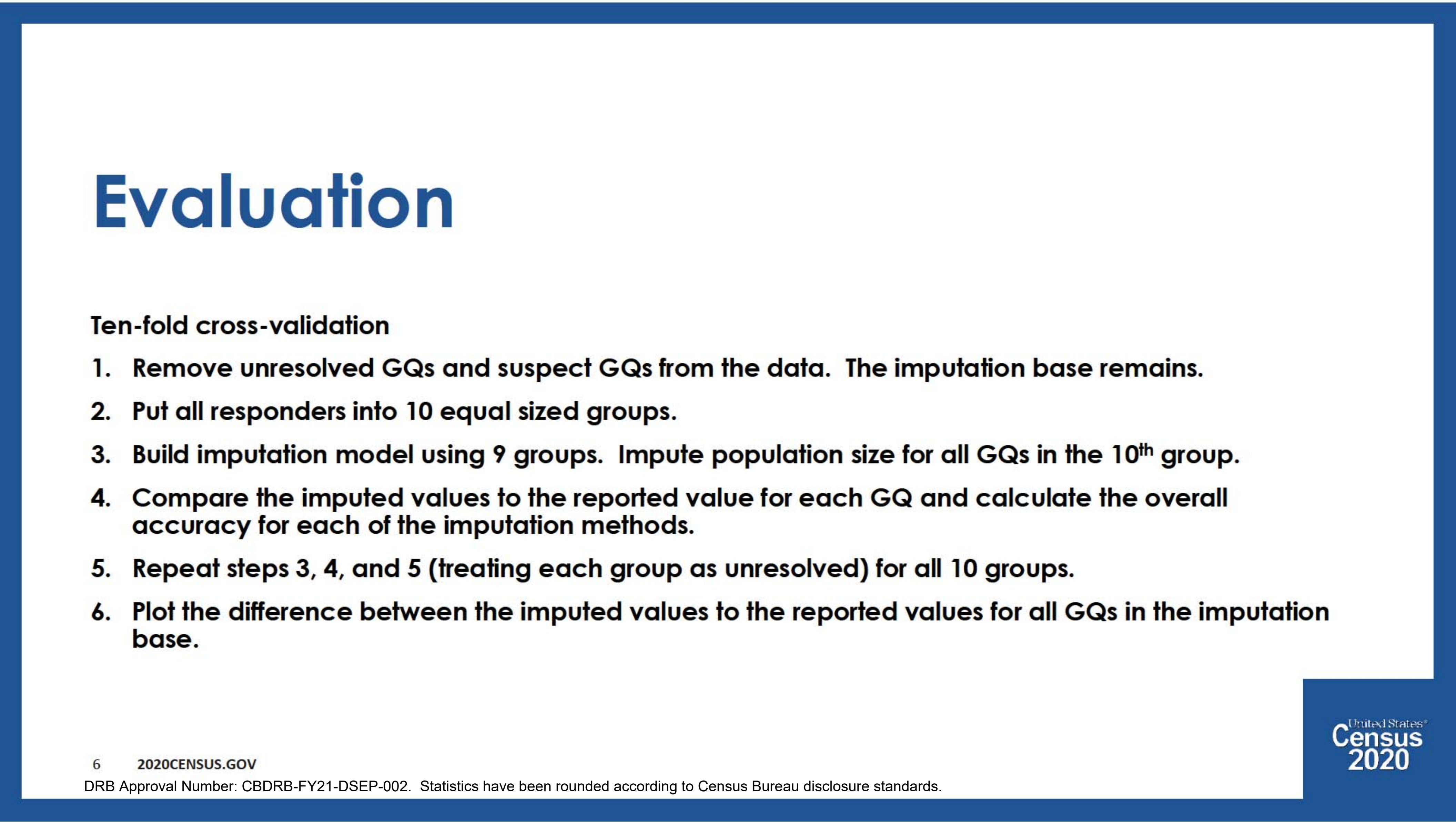
  Allocates reported facility population counts from Integrated Postsecondary Education Data System (IPEDS) to GQs

- **Poisson Regression**

  Like Logistic Regression, but better suited for count and rate data

- **Percentile Imputation**

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



Evaluation

Ten-fold cross-validation

1. **Remove unresolved GQs and suspect GQs from the data. The imputation base remains.**
2. **Put all responders into 10 equal sized groups.**
3. **Build imputation model using 9 groups. Impute population size for all GQs in the 10th group.**
4. **Compare the imputed values to the reported value for each GQ and calculate the overall accuracy for each of the imputation methods.**
5. **Repeat steps 3, 4, and 5 (treating each group as unresolved) for all 10 groups.**
6. **Plot the difference between the imputed values to the reported values for all GQs in the imputation base.**

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

[redacted]

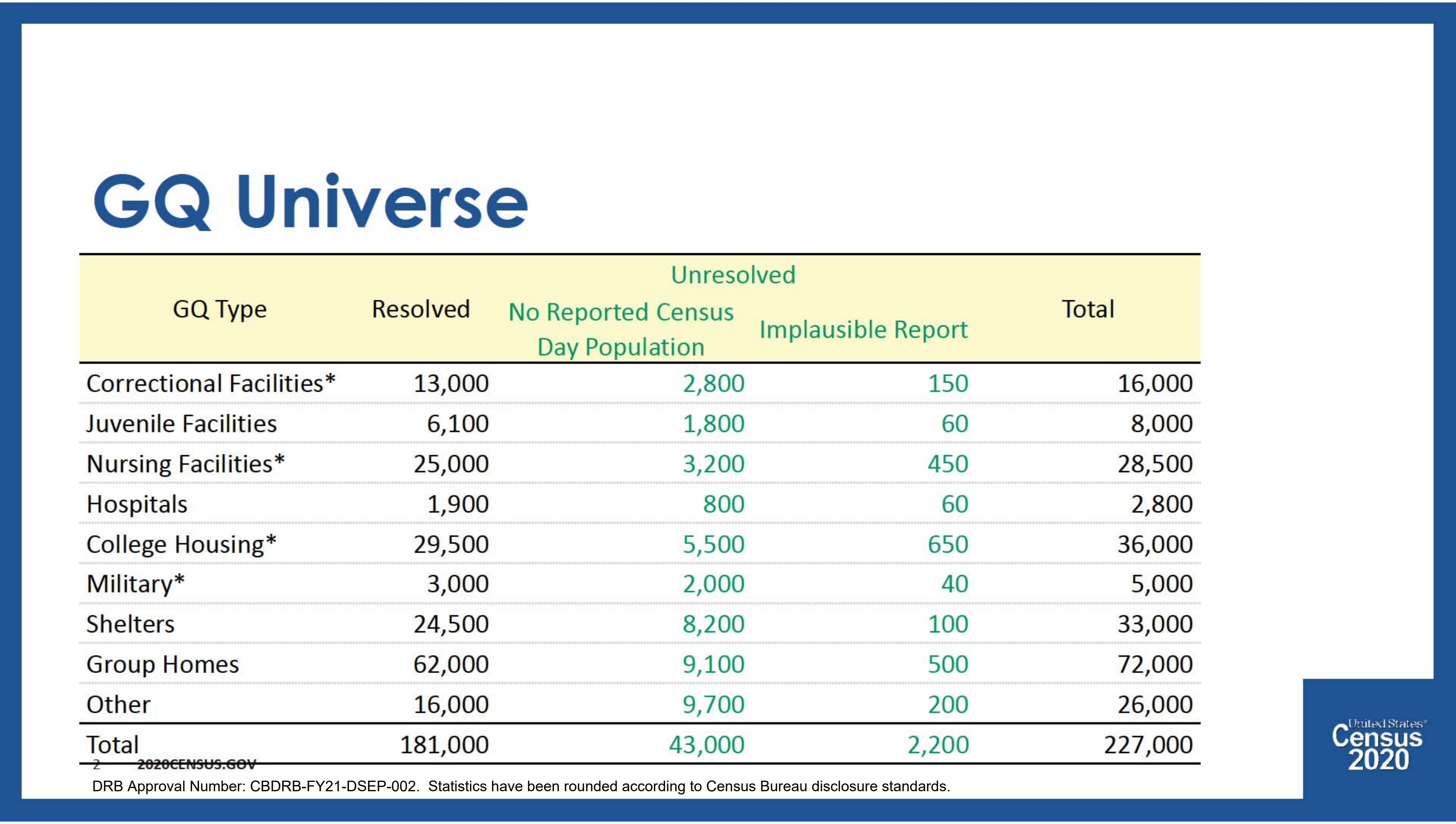
DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

[redacted]

Group Quarters Count Imputation Methodology

December 23, 2020

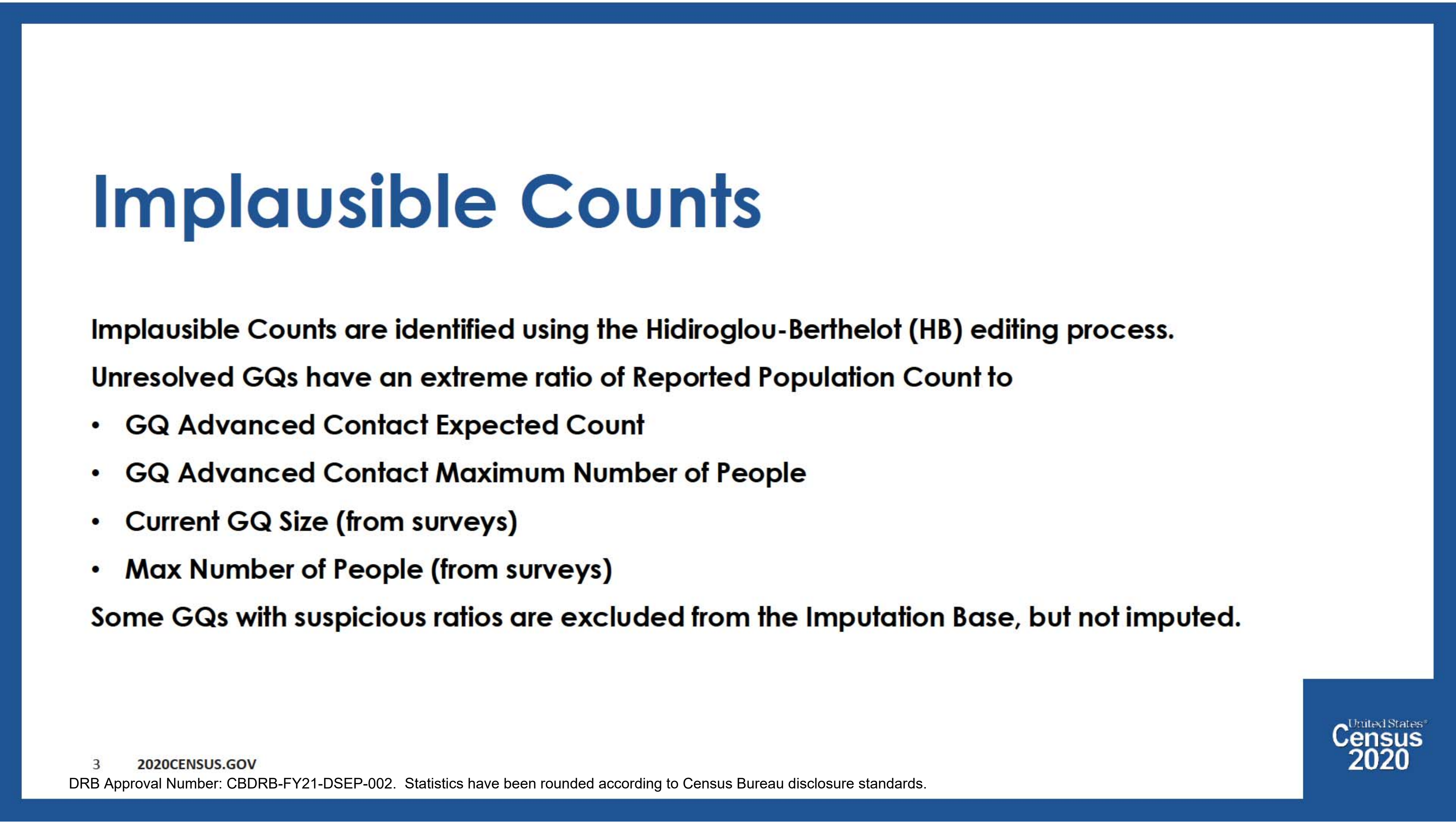
DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



GQ Universe

GQ Type	Resolved	Unresolved		Total
		No Reported Census Day Population	Implausible Report	
Correctional Facilities*	13,000	2,800	150	16,000
Juvenile Facilities	6,100	1,800	60	8,000
Nursing Facilities*	25,000	3,200	450	28,500
Hospitals	1,900	800	60	2,800
College Housing*	29,500	5,500	650	36,000
Military*	3,000	2,000	40	5,000
Shelters	24,500	8,200	100	33,000
Group Homes	62,000	9,100	500	72,000
Other	16,000	9,700	200	26,000
Total	181,000	43,000	2,200	227,000

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



Implausible Counts

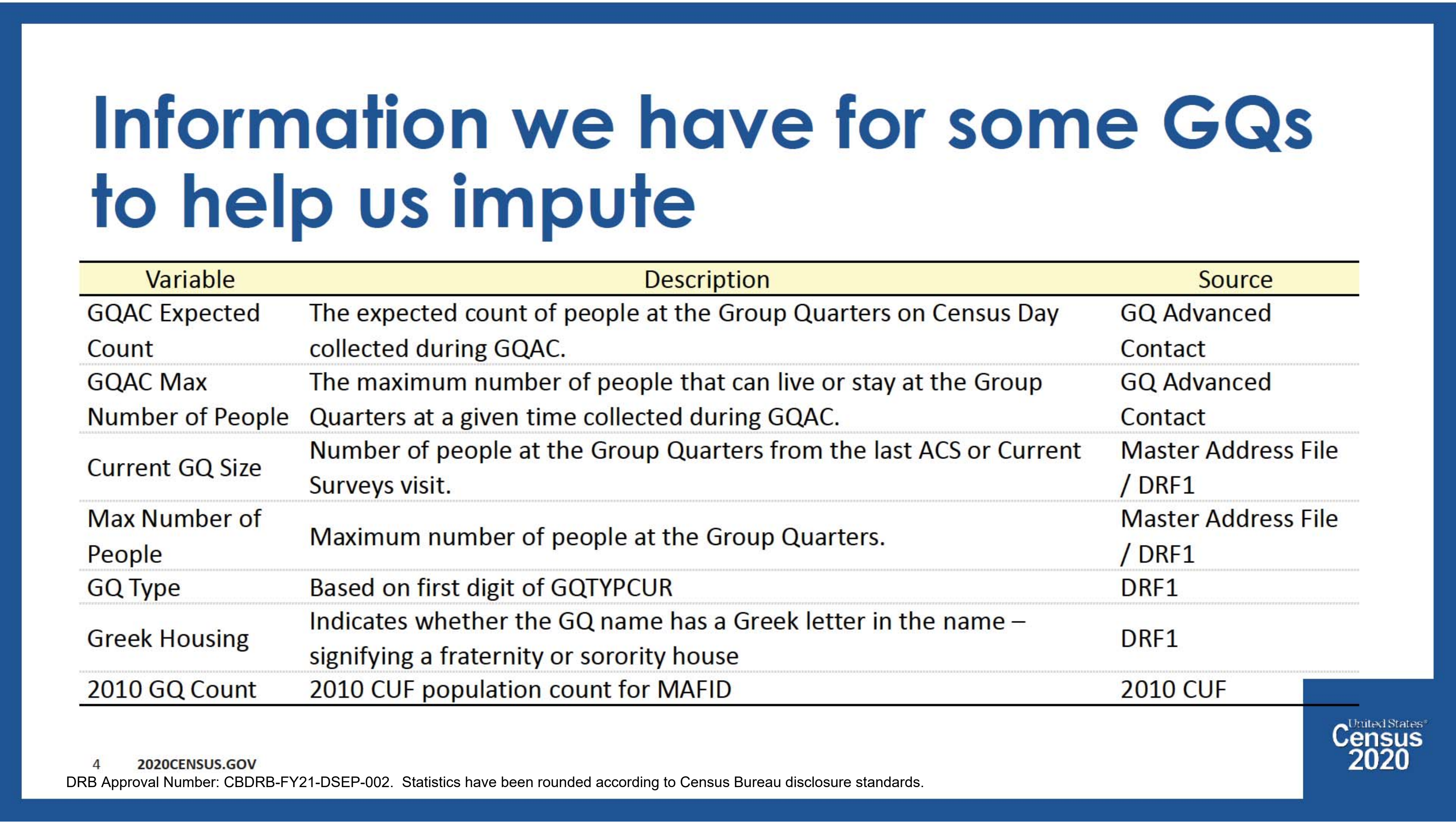
Implausible Counts are identified using the Hidiroglou-Berthelot (HB) editing process.

Unresolved GQs have an extreme ratio of Reported Population Count to

- GQ Advanced Contact Expected Count
- GQ Advanced Contact Maximum Number of People
- Current GQ Size (from surveys)
- Max Number of People (from surveys)

Some GQs with suspicious ratios are excluded from the Imputation Base, but not imputed.

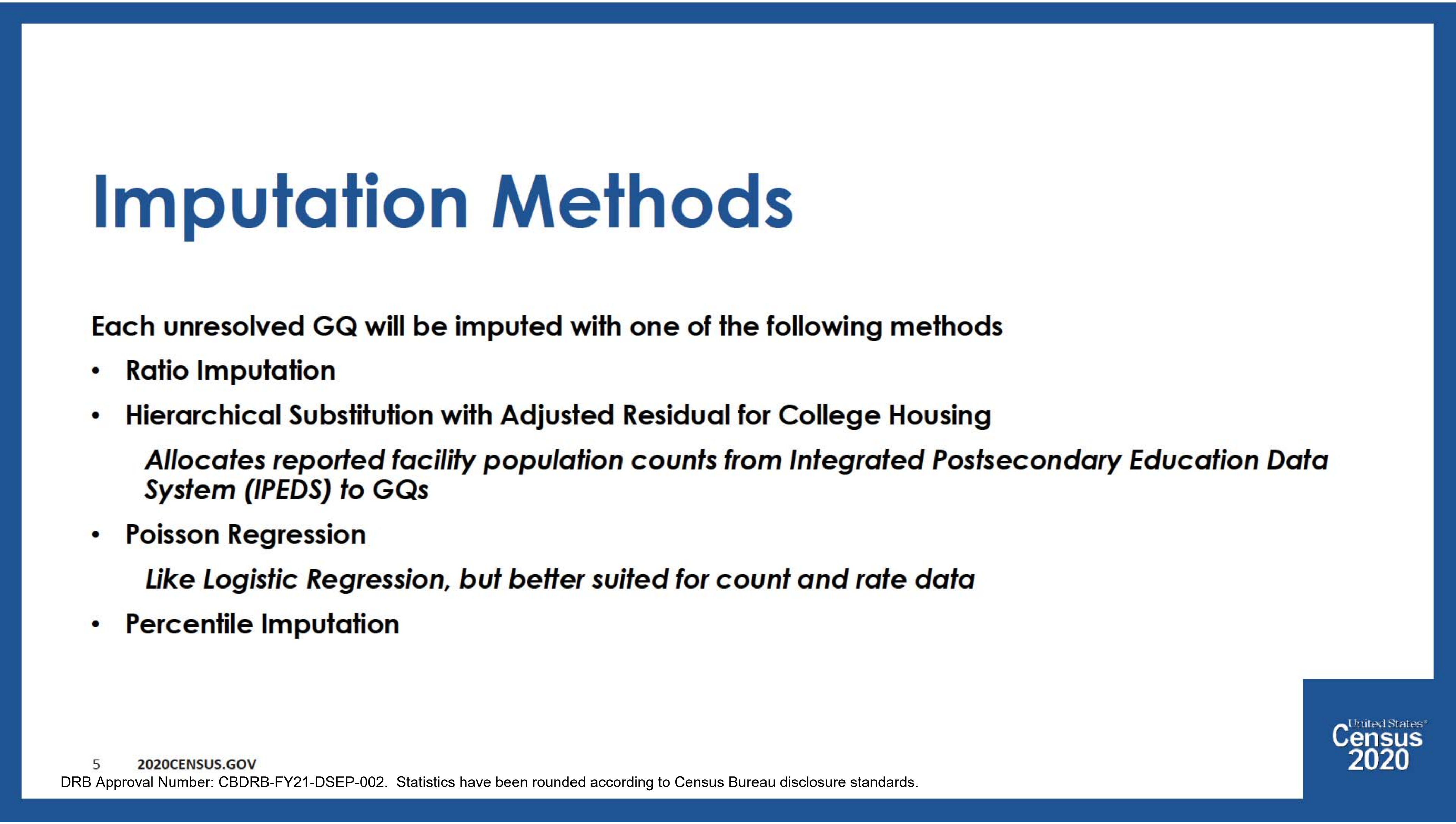
DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



Information we have for some GQs to help us impute

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



Imputation Methods

Each unresolved GQ will be imputed with one of the following methods

- **Ratio Imputation**
- **Hierarchical Substitution with Adjusted Residual for College Housing**

  Allocates reported facility population counts from Integrated Postsecondary Education Data System (IPEDS) to GQs

- **Poisson Regression**

  Like Logistic Regression, but better suited for count and rate data

- **Percentile Imputation**

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Evaluation

Ten-fold cross-validation

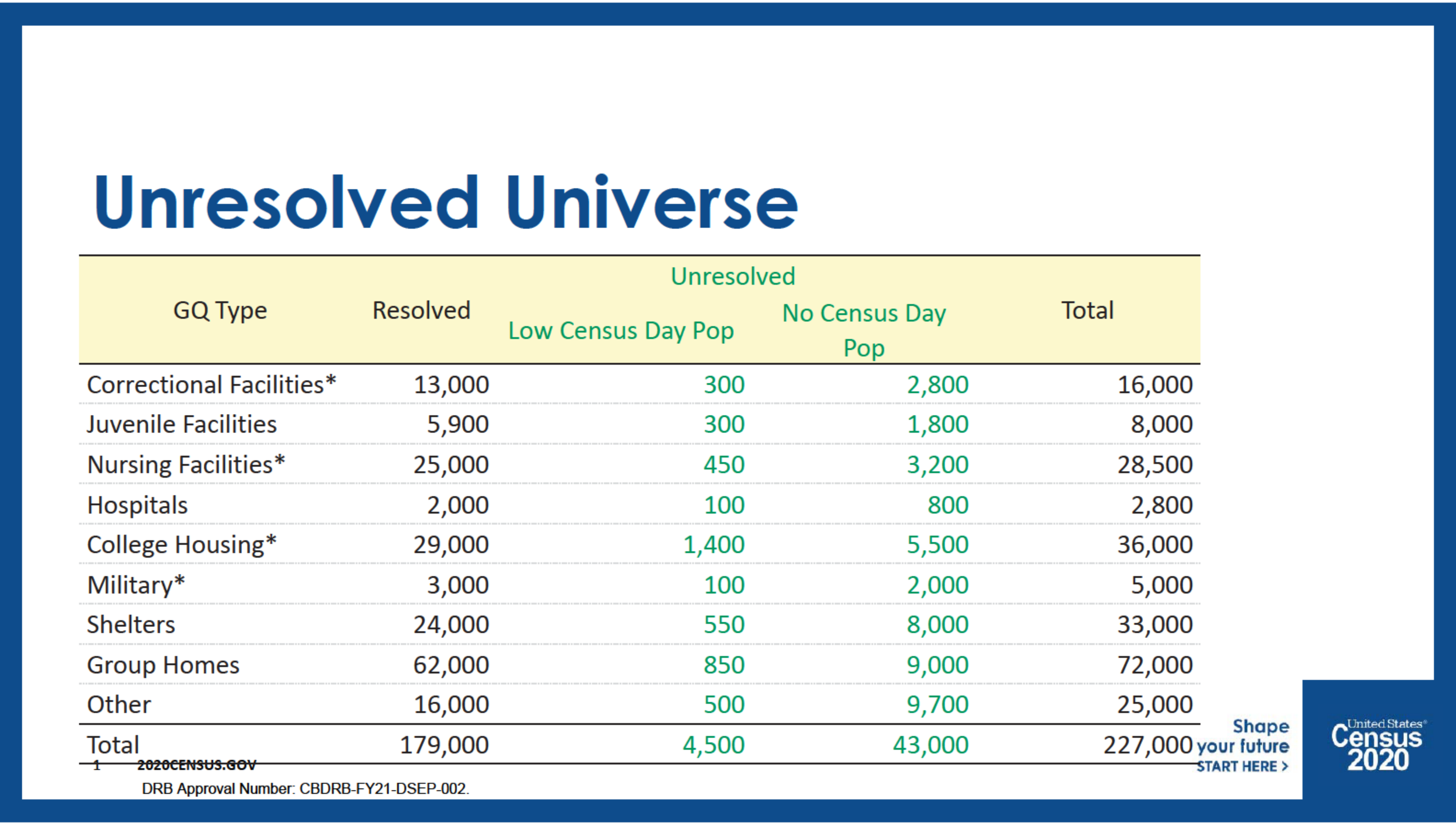
1. **Remove unresolved GQs and suspect GQs from the data. The imputation base remains.**
2. **Put all responders into 10 equal sized groups.**
3. **Build imputation model using 9 groups. Impute population size for all GQs in the 10th group.**
4. **Compare the imputed values to the reported value for each GQ and calculate the overall accuracy for each of the imputation methods.**
5. **Repeat steps 3, 4, and 5 (treating each group as unresolved) for all 10 groups.**
6. **Plot the difference between the imputed values to the reported values for all GQs in the imputation base.**

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

[redacted]



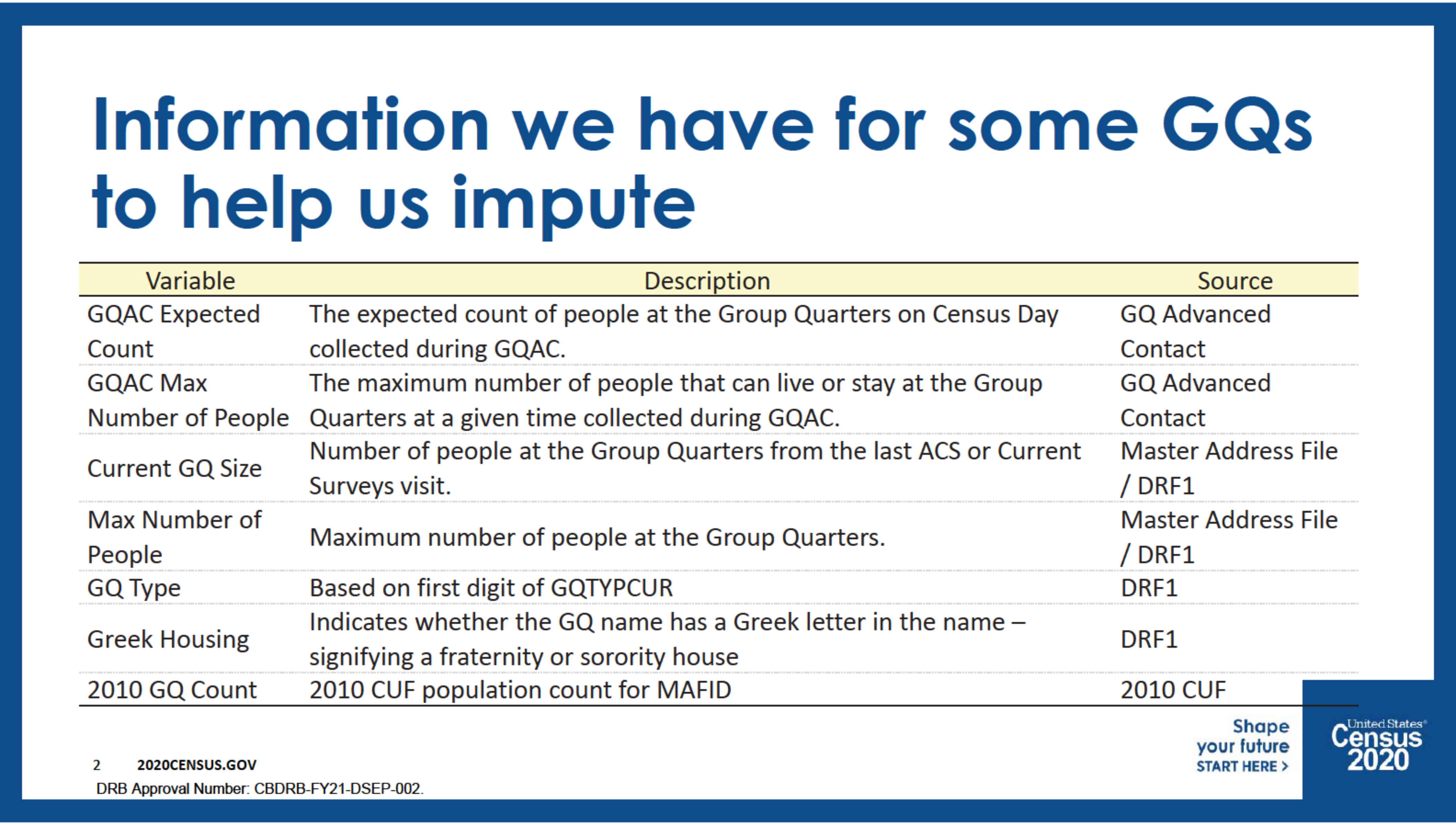
Unresolved Universe

GQ Type	Resolved	Unresolved		Total
		Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	13,000	300	2,800	16,000
Juvenile Facilities	5,900	300	1,800	8,000
Nursing Facilities*	25,000	450	3,200	28,500
Hospitals	2,000	100	800	2,800
College Housing*	29,000	1,400	5,500	36,000
Military*	3,000	100	2,000	5,000
Shelters	24,000	550	8,000	33,000
Group Homes	62,000	850	9,000	72,000
Other	16,000	500	9,700	25,000
Total	179,000	4,500	43,000	227,000

DRB Approval Number: CBDRB-FY21-DSEP-002.

Shape your future START HERE >

United States Census 2020



Information we have for some GQs to help us impute

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF

DRB Approval Number: CBDRB-FY21-DSEP-002.

Imputation Methods

Each unresolved GQ will be imputed with one of the following methods

- **Ratio Imputation**
- **Hierarchical Substitution with Adjusted Residual for College Housing**

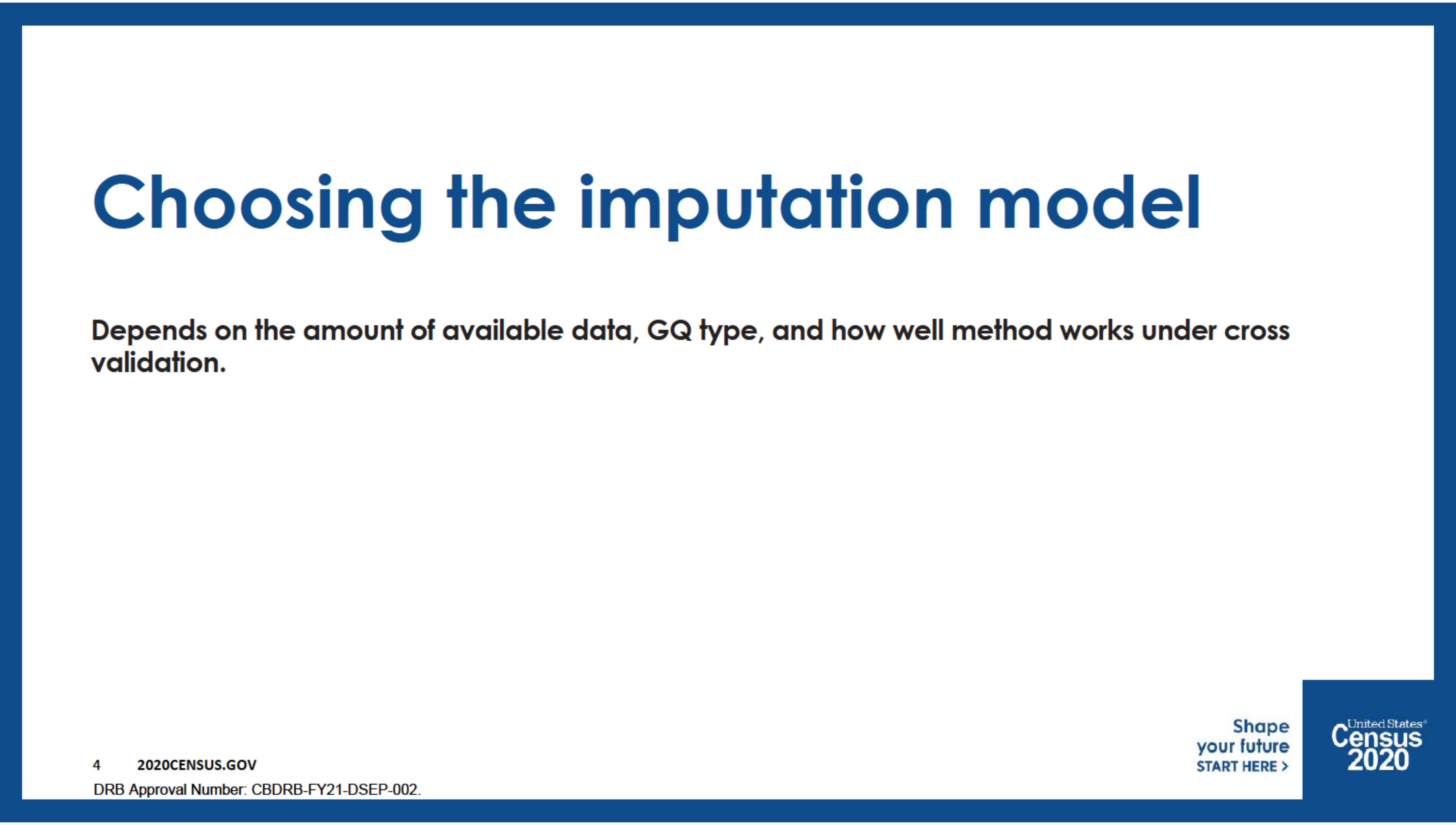
  Allocates reported facility population counts from Integrated Postsecondary Education Data System (IPEDS) to GQs

- **Poisson Regression**

  Like Logistic Regression, but better suited for count and rate data

- **Median Imputation**

DRB Approval Number: CBDRB-FY21-DSEP-002.



Choosing the imputation model

Depends on the amount of available data, GQ type, and how well method works under cross validation.

Evaluation

Ten-fold cross-validation

1. Remove unresolved GQs from the data.
2. Put all responders into 10 equal sized groups.
3. Build imputation model using 9 groups. Impute pop size for all GQs in the 10th group.
4. Compare the imputed values to the reported value for each GQ and calculate the overall accuracy for each of the four imputation methods.
5. Repeat steps 3, 4, and 5 (treating each group as unresolved) for all 10 groups.
6. Average the overall accuracy for all 10 groups.

DRB Approval Number: CBDRB-FY21-DSEP-002.

METHODOLOGY MEMO
Imputation Universe

- We won't impute for responses in the NPC work.
- Do we impute for all refusals?
- Do we just impute 0 pop counts?
- Do we impute cases with coutn discrepencies (however that is defined)?
- Do we assume all vacant discrepenay cases as occupied?
- Do we just impute when we don't have expected count?
- Which GQ types are in scope (501, 601 and 600, nursing homes OR all)

approach - different models for differnt types of GQs maybe different size

- Model building process (cross validation)
- What coviarates will we use
- different models for greek housing
- Make it clear that the approach is heirarchal
- For Poisson Regression what is the dependent variable and what is the offset (count / expected) OR (count / max)

Metrics to evaluate models

- Run model on half of data - Validation
- distance

Distributions

- what percent have expected pop count? What percent have max count? How many did we get advanced contact? How any have low ratio of people to expected count?
- How many "unresolved" GQs have ACS responses?
- How many GQs have 2010 data?
- How many GQs have some advance or preliminary counts?
- What variables are available?

Andrew Keller
Imputing GQ Pop Counts – Draft 1
December 6, 2020

Table 1: Input Data

	No Good Person	Has Good Person	Total
Occupied GQ	18,646	180,396	199,042
Delete GQ	7,225	381	7,606
Nonresidential GQ	2,373	76	2,449
Vacant During Visit, Open on Census Day	19,683	1,542	21,225
Refusal GQ	6,756	973	7,729
Vacant GQ	29,229	968	30,197
Total	83,912	184,336	268,248

1. Red and Green (223,163 cases)
    a. These are the resolved cases – use appropriate count
    b. Red are the donors on the models below
2. Blue (45,085 cases) – These are the unresolved cases. We believe them to be occupied, but do not have a good person count.

Business Rules

1. If the unresolved cases was a GQ in 2010 and had a pop count, I am going to directly assign that pop count.
2. If not, I use the modeled result.

Two Models

1. Has 2020 GQ Expected Count - Linear Regression Model
    a. DV: ratio of 2020 Good Person Count / 2020 GQ Expected Count
    b. 91,658 of the 180,396 cases have 2020 GQ Expected Count
    c. Score model over 9,020 unresolved cases with a 2020 GQ Expected Count. This outputs an estimated occupied ratio which I multiply by the 2020 GQ Expected count to get an imputed GQ count.
2. No 2020 GQ Expected Count - Linear Regression Model
    a. DV: 2020 Good Person Count
    b. 88,738 of the 180,396 cases without 2020 GQ Expected Count
    c. Score model over 36,065 unresolved cases without a 2020 GQ Expected Count. This outputs an imputed GQ count.

Results

1. Use 2020 ACS GQ Count As a Baseline
2. Compare Results Between No Imputation (Keeping a 0 for all Blue Cases) and Imputation (Applying Business Rules and Models

2020 ACS GQ Count – 8,084,362
2020 Census GQ Count (No Imputation) – 8,294,160
2020 Census GQ Count (With Imputation) – 10,198,552

Path	GQ	% of GQ	GQ People	% of GQ People
Resolved	223,163	83.2%	8,294,160	81.3%
Has 2020 Expected Pop, Use 2010 GQ Count	5,650	2.1%	252,257	2.5%
Has 2020 Expected Pop, Use Model	3,370	1.3%	300,883	3.0%
Without 2020 Expected Pop, Use 2010 GQ Count	11,393	4.2%	462,162	4.5%
Without 2020 Expected Pop, Use Model	24,672	9.2%	889,090	8.7%
Total	268,248	100.0%	10,198,552	100.0%

12/6/20 – Models being refined

2020 GQ Count (No Imputation Ratio) – 1.03
2020 GQ Count (With Imputation) – 1.26

Model Appendix

1. Has 2020 GQ Expected Count

```
proc reg data=yesmaxmod outest=yesmaxparam;
     model filledratio = /* feddc */ statejail localjail housejail nursing
college military homeless soup /* uaa */ group dne2010 ar1 ar2 ar3 ar6 max5l
max1\
00m nomax;
run;
```

2. No 2020 GQ Expected Count

```
proc reg data=nomaxmod outest=nomaxparam;
     model gp = feddc statejail  localjail housejail nursing college military
homeless soup uaa group dne2010 ar1 ar2 ar3 ar6 max5l max100m nomax;
run;
```

County Distribution of 2020 Census / 2020 ACS - GQ Person Ratios Before Imputation

[redacted]

County Distribution of 2020 Census / 2010 ACS - GQ Person Ratios After Imputation

[redacted]

Andrew Keller
Imputing GQ Pop Counts – Draft 1
December 13, 2020

New Input File: GQ_MAFID_CNTS_drf2_cdl_121320.csv

Table 1: Input Data

GQ Status	No Good Person (GP)	Has Good Person	Total
Occupied GQ	17,000	181,000	197,000
Delete GQ	7,200	450	7,600
Nonresidential GQ	2,400	100	2,500
Vacant During Visit, Open on Census Day	19,500	1,900	21,500
Refusal GQ	6,700	1,100	7,800
Vacant GQ	29,000	1,100	30,500
Total	82,000	185,000	267,000

To determine the GQ status: use FOCS_ER_CB_CODE

To determine the GQ has good persons (and the GQ count), I use the gp value, but I overwrite with this logic.

```
    if gp_psa > 0 then gp = gp_psa;
    if gp = . and ddp = (0,.) then gp = cdlper;

    if gp > 0 then gpy = 1; else gpy = 0;
```

To determine the unresolved cases:

```
unres = 0;
if FOCS_ER_CB_CODE in ('','O','R') and gpy = 0 then unres = 1;
```

1. Red and Green (224,000 cases)
    a. These are the resolved cases – use appropriate count
    b. Red are the donors on the models below
2. Blue (43,000 cases) – These are the **unresolved** cases. We believe them to be occupied, but do not have a good person count.

Hierarchical Approach

1. CES 501 Approach (not in this simulation)
2. Ratio-Adjusted GQ Advanced Contact (AC) Imputation – Given we have an GQAC count, calculate a Good Person / GQAC Expected ratio at the following levels and multiply by the GQAC Expected count
    a. Nest on State and GQ Type
    b. Nest on National GQ Type
    c. National
3. Poisson Regression Model using Current Surveys GQ count as offset – Given we have a current surveys GQ count, fit a model
4. Mean GP Imputation – Given we don't have an GQAC or Current Surveys GQ count, calculate a mean GP count
    a. Nest on State and GQ Type
    b. Nest on National GQ Type
    c. National

Results

1. Use 2020 ACS GQ Count As a Baseline
2. Compare Results Between No Imputation (Keeping a 0 for all Blue Cases) and Imputation (Applying Business Rules and Models

2020 ACS GQ Count – 8,084,000
2020 Census GQ Count (No Imputation) – 8,122,000
2020 Census GQ Count (With Imputation) – 9,842,000

Path	GQ	% of GQ	GQ People	% of GQ People
Resolved	224,000	83.9%	8,122,000	82.5%
CES 501s				
Ratio-Adjusted GQAC From State x GQTYP	8,600	3.2%	393,000	4.0%
Ratio-Adjusted GQAC From GQTYP	(D)	(D)	5,300	0.1%
Ratio-Adjusted GQAC From National	N<15	N<15	900	0.0%
Model	9,300	3.5%	355,000	3.6%
Good Person Mean From State x GQTYP	25,000	9.4%	936,000	9.5%
Good Person Mean From GQTYP	400	0.1%	29500	0.3%
Total	267,000	100.0%	9,842,000	100.0%

Of the 181,000 occupied cases with Good Person count, **92,000** have an expected count.

```
                                               Cumulative    Cumulative
hasexp    unres    Frequency     Percent        Frequency       Percent
-----------------------------------------------------------------------
     0        0        88500        44.8            88500          44.8
     0        1        16000         8.10          104000          52.6
     1        0        92000        46.58          196000          99.2
     1        1         1000         0.50          197000         100.00
```

Look at the ratio of good person count to expected count among the resolved cases summed nationally. About 79.7% of the expected count shows up in the good person count.

```
Obs    expratio

 1      0.7974
```

Do Same Analysis By GQ Type

```
                           expratio_
Obs          GQTYPCUR       gqtypcur

  1               103       0.8038
  2               104       0.7075
  3               105       0.7354
  4               106       1.157
  5               201       0.7416
  6               202       0.6838
  7               203       0.5769
  8               301       0.8668
  9               401       0.8475
 10               402       0.7996
 11               403       0.7102
 12               404       0.2182
 13               405       0.7498
 14               501       0.7938
```

For nearly all states, a ratio can by computed for the GQ Types

About 20% of the unresolved universe can be covered in this manner.

```
                                               Cumulative    Cumulative
```

```
hasexp    Frequency     Percent     Frequency      Percent
-----------------------------------------------------------
     0       34500       80.05         34500        80.05
     1        8600       19.95         43100       100.00
```

Apply the .

d. Apply Business Rules

i. If 2010 GQ or Occ HU – directly insert count

ii. Go to Model if no 2010 GQ or Occ HU value

e. Dependent variable is Good Person Count

f. New Model for Each Major GQ Type

g. Independent variables

i. Didn't Exist in 2010

ii. Occ HU in 2010

iii. 2010 GQ count 0-19,20-49,50-99,100-199,200+

iv. AR count of 1,2,3-5,6+

v. NO 2020 GQ Max, GQ Max < 5, 6-100, 100+

vi. NO 2020 GQ Exp, GQ Exp < 5, 6-100, 100+

vii. Greek?

viii. Dorm?, Hall?, Housing?, College?

ix. Dorm? Interacted with NO 2020 GQ Max

x. Dorm? Interacted with NO 2020 GQ Exp

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Andrew Keller
Imputing GQ Pop Counts – Draft 1
December 14, 2020

New Input File: GQ_MAFID_CNTS_drf2_cdl_121320.csv

Table 1: Input Data

GQ Status	No Good Person (GP)	Has Good Person	Total
Occupied GQ	17,000	181,000	197,000
Delete GQ	7,200	450	7,600
Nonresidential GQ	2,400	100	2,500
Vacant During Visit, Open on Census Day	19,500	1,900	21,500
Refusal GQ	6,700	1,100	7,800
Vacant GQ	29,000	1,100	30,500
Total	82,000	185,000	267,000

To determine the GQ status: use FOCS_ER_CB_CODE

To determine the GQ has good persons (and the GQ count), I use the gp value, but I overwrite with this logic.

```
    if gp_psa > 0 then gp = gp_psa;
    if gp = . and ddp = (0,.) then gp = cdlper;

    if gp > 0 then gpy = 1; else gpy = 0;
```

To determine the unresolved cases:

```
unres = 0;
if FOCS_ER_CB_CODE in ('','O','R') and gpy = 0 then unres = 1;
```

1. Red and Green (224,000 cases)
    a. These are the resolved cases – use appropriate count
    b. Red are the donors on the models below
2. Blue (43,000 cases) – These are the **unresolved** cases. We believe them to be occupied, but do not have a good person count.

Hierarchical Approach

1. CES 501 Approach (not in this simulation)
2. Ratio-Adjusted GQ Advanced Contact (AC) Imputation – Given we have an GQAC count, calculate a Good Person / GQAC Expected ratio at the following levels and multiply by the GQAC Expected count
    a. Nest on State and GQ Type
    b. Nest on National GQ Type
    c. National
3. Poisson Regression Model using Current Surveys GQ count as offset – Given we have a current surveys GQ count, fit a model (see appendix for vars)
4. Mean GP Imputation – Given we don't have an GQAC or Current Surveys GQ count, calculate a mean GP count
    a. Nest on State and GQ Type
    b. Nest on National GQ Type
    c. National

Results

1. Use 2020 ACS GQ Count As a Baseline
2. Compare Results Between No Imputation (Keeping a 0 for all Blue Cases) and Imputation (Applying Business Rules and Models

2020 ACS GQ Count – 8,084,000
2020 Census GQ Count (No Imputation) – 8,122,000
2020 Census GQ Count (With Imputation) – 9,842,000

Path	GQ	% of GQ	GQ People	% of GQ People
Resolved	224,000	83.9%	8,122,000	82.5%
CES 501s				
Ratio-Adjusted GQAC From State x GQTYP	8,600	3.2%	393,000	4.0%
Ratio-Adjusted GQAC From GQTYP	(D)	(D)	5,300	0.1%
Ratio-Adjusted GQAC From National	N<15	N<15	900	0.0%
Model	9,300	3.5%	355,000	3.6%
Good Person Mean From State x GQTYP	25,000	9.4%	936,000	9.5%
Good Person Mean From GQTYP	400	0.1%	29500	0.3%
Total	267,000	100.0%	9,842,000	100.0%

Breakout of Good Person Mean From State x GQTYP

Path	GQ	% of GQ	GQ People	% of GQ People
Correctional Facilities*	1,600	6.4%	220,000	23.5%
Juvenile Facilities	650	2.6%	9,300	1.0%
Nursing Facilities*	1,200	4.8%	69,500	7.4%
Hospitals	350	1.4%	15,500	1.7%
College Housing*	2,500	10.0%	239,000	25.5%
Military*	700	2.8%	65,000	6.9%
Shelters	6,500	26.0%	158,000	16.9%
Group Homes	3,900	15.6%	45,500	4.9%
Other	7,300	29.2%	114,000	12.2%
Total	25,000	100.0%	936,000	100.0%

12/15 Addendum:

A. Applying 65th percentile value instead of taking means
B. Taking 40 (fit) / 60 (score) sample of data, look at bias (imputed pop – true pop) measure over 10 simulations:

```
Obs    _TYPE_    _FREQ_    Bias    SE(Bias)
 1        0        10      9800    65400
```

County Distribution of 2020 Census / 2020 ACS - GQ Person Ratios Before Imputation

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

County Distribution of 2020 Census / 2010 ACS - GQ Person Ratios After Imputation

[redacted]

Appendix

Poisson Regression Model with Offset

d. Dependent variable is Good Person Count
e. New Model for Each Major GQ Type
f. Independent variables
    i. Didn't Exist in 2010
    ii. Occ HU in 2010
    iii. 2010 GQ count 0-19,20-49,50-99,100-199,200+
    iv. AR count of 1,2,3-5,6+
    v. NO 2020 GQAC Max, GQAC Max < 5, 6-100, 100+
    vi. NO 2020 GQAC Exp, GQAC Exp < 5, 6-100, 100+
    vii. Greek?
    viii. Dorm?, Hall?, Housing?, College?
    ix. Dorm? Interacted with NO 2020 GQ Max
    x. Dorm? Interacted with NO 2020 GQ Exp

Andrew Keller
Imputing GQ Pop Counts – Draft 1
December 16, 2020

New Input File: GQ_MAFID_CNTS_drf2_cdl_121320.csv

Table 1: Input Data

GQ Status	No Good Person (GP)	Has Good Person	Total
Occupied GQ	17,000	181,000	197,000
Delete GQ	7,200	450	7,600
Nonresidential GQ	2,400	100	2,500
Vacant During Visit, Open on Census Day	19,500	1,900	21,500
Refusal GQ	6,700	1,100	7,800
Vacant GQ	29,000	1,100	30,500
Total	82,000	185,000	267,000

To determine the GQ status: use FOCS_ER_CB_CODE

To determine the GQ has good persons (and the GQ count), I use the gp value, but I overwrite with this logic.

```
    if gp_psa > 0 then gp = gp_psa;
    if gp = . and ddp = (0,.) then gp = cdlper;

    if gp > 0 then gpy = 1; else gpy = 0;
```

To determine the unresolved cases:

```
unres = 0;
if FOCS_ER_CB_CODE in ('','O','R') and gpy = 0 then unres = 1;
```

1. Red and Green (224,000 cases)
    a. These are the resolved cases – use appropriate count
    b. Red are the donors on the models below
2. Blue (43,000 cases) – These are the **unresolved** cases. We believe them to be occupied, but do not have a good person count.

Hierarchical Approach

1. CES 501 Approach (not in this simulation)
2. Ratio-Adjusted GQ Advanced Contact (AC) Imputation – Given we have an GQAC count, calculate a Good Person / GQAC Expected ratio at the following levels and multiply by the GQAC Expected count
    a. Nest on State and GQ Type
    b. Nest on National GQ Type
    c. National
3. Poisson Regression Model using Current Surveys GQ count as offset – Given we have a current surveys GQ count, fit a model (see appendix for vars)
4. Mean GP Imputation – Given we don't have an GQAC or Current Surveys GQ count, calculate a mean GP count
    a. Nest on State and GQ Type
    b. Nest on National GQ Type
    c. National

Results

1. Use 2020 ACS GQ Count As a Baseline
2. Compare Results Between No Imputation (Keeping a 0 for all Blue Cases) and Imputation (Applying Business Rules and Models

2020 ACS GQ Count – 8,084,000
2020 Census GQ Count (No Imputation) – 8,122,000
2020 Census GQ Count (With Imputation) – 9,332,000

Path ID	Path	GQ	% of GQ	GQ People	% of GQ People
100	Resolved	224,000	83.9%	8,122,000	87.0%
310	Ratio-Adjusted GQAC From State x GQTYP	8,600	3.2%	393,000	4.2%
320	Ratio-Adjusted GQAC From GQTYP	(D)	(D)	5,300	0.1%
330	Ratio-Adjusted GQAC From National	N<15	N<15	900	0.0%
410	Model	9,300	3.5%	355,000	3.8%
510	2010-Adjusted From State x GQTYP	4,200	1.6%	37,000	0.4%
520	2010-Adjusted GQAC From GQTYP	50	0.0%	1,100	0.0%
610	Good Person 70th percentile From State x GQTYP	20,500	7.7%	402,000	4.3%
620	Good Person 70th percentile From GQTYP	400	0.1%	15,500	0.2%
	Total	267,000	100.0%	9,332,000	100.0%

Applying Models to 12/16 Truth Decks

Bias = SUM(Imputed GQ Pop) – SUM(Provided GQ Pop)

National

Path ID	Path	Mean(Bias)	SE(Bias)
National	National	9,900	51,620

Path-Level

Path ID	Path	Mean(Bias)	SE(Bias)
310	Ratio-Adjusted GQAC From State x GQTYP	21,000	39,400
320	Ratio-Adjusted GQAC From GQTYP	1,500	1,064
410	Model	-28,500	15,290
510	2010-Adjusted From State x GQTYP	-13,000	3,720
520	2010-Adjusted GQAC From GQTYP	30	47
610	Good Person 65th Percentile From State x GQTYP	26,000	8,856
620	Good Person 65th Percentile From GQTYP	2,900	1,678

GQ Type-Level

GQ ID	GQ Type	Mean(Bias)	SE(Bias)
104	Local Jails and Other Municipal Confinement Facilities	-5,200	4,003
105	Correction Residential Facilities	1,600	1,088
106	Military Disciplinary Barracks and Jails	200	174
201	Group Homes for Juveniles (non-correctional)	30	607
202	Residential Treatment Centers for Juveniles (non-correctional)	600	456
203	Correctional Facilities Intended for Juveniles (training schoo	450	649
301	Nursing Facilities/Skilled-Nursing Facilities	51,500	5,854
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	400	1,510
402	Hospitals with Patients Who Have No Usual Home Elsewhere	650	1,061

403	In-Patient Hospice Facilities	-40	370
404	Military Treatment Facilities with Assigned Active Duty Patien	-40	318
405	Residential Schools for People with Disabilities	1,100	1,027
501	College/University Student Housing	-36,500	15,930
601	Military Quarters	3,800	5,923
602	Military Ships	-2,900	994
701	Emergency and transitional shelters (with sleeping facilities)	-6,400	2,130
702	Soup Kitchens	-1,400	3,948
704	Domestic Violence Shelters	-750	751
706	Regularly Scheduled Mobile Food Vans	-12,500	1,771
801	Targeted Non-Sheltered Outdoor Locations	-25,500	3,089
802	Residential Treatment Centers for Adults (non-correctional)	-3,500	1,823
900	Maritime/Merchant Vessels	600	150
901	Workers' Group Living Quarters and Job Corps Centers	-5,200	2,957
903	Religious Group Quarters (convents, monasteries, abbeys)	N<15	N<15
999	Living Quarters for Victims of Natural Disasters	-2,300	631

Selected GQ Type By Path 201 – Nursing Facilities/Skilled-Nursing Facilities

All	Nursing Facilities/Skilled-Nursing Facilities	51,500	5,854
610	Good Person 65th Percentile From State x GQTYP	58,000	2,379
410	Model	-200	842
310	Ratio-Adjusted GQAC From State x GQTYP	-2,500	5,565
510	2010-Adjusted From State x GQTYP	-3,500	471

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

<u>Appendix</u>

Poisson Regression Model with Offset

d. Dependent variable is Good Person Count
e. New Model for Each Major GQ Type
f. Independent variables
    i. Didn't Exist in 2010
    ii. Occ HU in 2010
    iii. 2010 GQ count 0-19,20-49,50-99,100-199,200+
    iv. AR count of 1,2,3-5,6+
    v. NO 2020 GQAC Max, GQAC Max < 5, 6-100, 100+
    vi. NO 2020 GQAC Exp, GQAC Exp < 5, 6-100, 100+
    vii. Greek?
    viii. Dorm?, Hall?, Housing?, College?
    ix. Dorm? Interacted with NO 2020 GQ Max
    x. Dorm? Interacted with NO 2020 GQ Exp

Andrew Keller
Imputing GQ Pop Counts
December 17, 2020

Input Data – Does not incorporate Call-in or web scraping results

Table 1: Input Data

GQ Status	No Good Person (GP)	Has Good Person	Total
Occupied GQ	17,000	181,000	197,000
Delete GQ	7,200	450	7,600
Nonresidential GQ	2,400	100	2,500
Vacant During Visit, Open on Census Day	19,500	1,900	21,500
Refusal GQ	6,700	1,100	7,800
Vacant GQ	29,000	1,100	30,500
Total	82,000	185,000	267,000

Make 10 replicates each of 10% missingness

Four Methods:

1. Ratio-Adjusted 2020 GQ Advanced Contact (AC) Expected Count – Needs GQ AC count
   a. If we have 2020 GQAC Expected count, we adjust it by a ratio determined by Good Person / GQAC ratio within state and GQ Type

2. Poisson Model (Applied to Only 101,103,104,301,501,601) – Needs GQ Current MAX Size
   a. Model Count offset GQ Current Max Size

3. Ratio-Adjusted 2010 Census GQ Count – Needs 2010 Census GQ count
   a. If we have 2010 Census count, we adjust it by a ratio determined by Good Person / 2010 Census ratio within state and GQ Type

4. Take 65th percentile good person count with state and GQ type as imputed count – Needs nothing

Given We Can Apply All 4 methods, which is best?

Applying Models to 12/17 Truth Decks

Bias = SUM(Imputed GQ Pop) – SUM(Provided GQ Pop)

GQ Type-Level

GQ ID	GQ Type	Mean # of GQs	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
104	Local Jails and Other Municipal Confinement Facilities	150	2,900	3,168	N<15	378	2,800	1,318	9,300	1,073
301	Nursing Facilities/Skilled-Nursing Facilities	1,300	1,300	979	-1,100	1,047	6,000	842	6,800	1,425
501	College	1,200	8,600	4,181	100	1,701	11,000	2,841	16,000	3,709
601	Military	70	1,000	859	100	1,139	1,400	1,040	2,000	1,378

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Andrew Keller
Imputing GQ Pop Counts
December 20, 2020

Input Data – Does not incorporate Call-in or web scraping results

Table 1: Input Data

GQ Status	No Good Person (GP)	Has Good Person	Total
Occupied GQ	17,000	181,000	197,000
Delete GQ	7,200	450	7,600
Nonresidential GQ	2,400	100	2,500
Vacant During Visit, Open on Census Day	19,500	1,900	21,500
Refusal GQ	6,700	1,100	7,800
Vacant GQ	29,000	1,100	30,500
Total	82,000	185,000	267,000

Make 10 replicates each of 10% missingness – using Juli's indicators for cases that should be suppressed or imputed.

Test 1: We have a positive count for ALL 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

Four Methods:

1. Ratio-Adjusted 2020 GQ Advanced Contact (AC) Expected Count – Needs GQ AC count
    a. If we have 2020 GQAC Expected count, we adjust it by a ratio determined by Good Person / GQAC ratio within state and GQ Type

2. Poisson Model –
    a. Model Count offset GQ Current Max Size

```
proc genmod data = nomaxmod;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

3. Take variable percentile good person count with state and GQ type as imputed count – Needs nothing
    a. 104, 801, 802, 901 – use 70th
    b. 301 – use 55th
    c. 501 – use 68th
    d. All others – use 65th
4. CES IPEDS method

Test 2: We have a positive count for AT LEAST ONE of 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

Three Methods:

1. Ratio-adjustment with hierarchy:
    a. Ratio-Adjusted 2020 GQ Advanced Contact (AC) Expected Count – Needs GQ AC count

b. Ratio-Adjusted 2020 GQAC Max Size – Needs GQ AC Max Size
c. Ratio-Adjusted Current Surveys Count – Needs Current Survey count
d. Ratio-Adjusted Current Surveys Max Size – Needs Current Survey Max Size

2. Take variable percentile good person count with state and GQ type as imputed count – Needs nothing
3. CES IPEDS method

Test 3: We nothing from 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

Two Methods:

1. Take variable percentile good person count with state and GQ type as imputed count – Needs nothing
2. CES IPEDS method

Results

Applying Models to 12/17 Truth Decks
Bias = SUM(Imputed GQ Pop) – SUM(Provided GQ Pop)

Test 1: We have a positive count for ALL 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

GQ Type-Level

GQ ID	GQ Type PUT IN MEDIAN BIAS	Mean # of GQs	Ratio Adjusted		Poisson		Median		IPEDS	
			Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
US	National	6,500	3,700	3,110	3,400	3,356	-6,900	11,920		
104	Local Jails and Other Municipal Confinement Facilities	200	400	3,152	100	2,724	-1,200	3,456		
105	Correctional Residential Facilities	30	20	396	20	485	-150	671		
106	Military Disciplinary Barracks and Jails	N<15	N<15	N<15	N<15	N<15	N<15	N<15		
201	Group Homes for Juveniles (non-correctional)	150	90	96	80	135	-300	157		
202	Residential Treatment Centers for Juveniles (non-correctional)	90	60	137	40	153	-150	191		
203	Correctional Facilities Intended for Juveniles (training schoo	60	100	426	30	233	-50	189		
301	Nursing Facilities/Skilled-Nursing Facilities	1,400	750	1,300	750	1,715	-1,200	2,265		
301	Nursing Facilities/Skilled-Nursing Facilities (Total Number of Occupied Beds)	700	150	694	300	1,025	-6,500	1,681	-3,600	585
301	Nursing Facilities/Skilled-Nursing Facilities (Number of All Beds)	700	150	694	300	1,025	-6,500	1,681	16,500	1,201
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	30	90	435	N<15	N<15	-200	1,081		
402	Hospitals with Patients Who Have No Usual Home Elsewhere	N<15	30	62	N<15	N<15	40	171		
403	In-Patient Hospice Facilities	20	30	50	N<15	N<15	-20	102		
404	Military Treatment Facilities with Assigned Active Duty Patien	N<15	N<15	N<15	N<15	N<15	N<15	N<15		

405	Residential Schools for People with Disabilities	20	70	104	N<15	N<15	-20	400		
501	College/University Student Housing	1,300	600	2,244	700	2,520	2,100	9,474		
501	College/University Student Housing – universe reduced for IPEDS	1,300	600	2,244	700	2,520	2,100	9,474	8,200	1,726
601	Military Quarters	80	-60	1,315	N<15	N<15	-60	2,229		
701	Emergency and transitional shelters (with sleeping facilities)	150	150	429	80	407	-750	733		
702	Soup Kitchens	70	80	605	150	588	-100	948		
704	Regularly Scheduled Mobile Food Vans	N<15	N<15	N<15	N<15	N<15	-20	36		
706	Targeted Non-Sheltered Outdoor Locations	20	20	102	N<15	N<15	-30	119		
801	Group Homes Intended for Adults (non-correctional)	2,100	850	319	1,100	658	-2,600	487		
802	Residential Treatment Centers for Adults (non-correctional)	350	200	217	200	199	-800	463		
901	Workers' Group Living Quarters and Job Corps Centers	250	150	250	100	248	-750	739		
903	Living Quarters for Victims of Natural Disasters	N<15	N<15	N<15	N<15	N<15	20	106		
999	Other	N<15	N<15	53	N<15	N<15	-20	39		

Test 2: We have a positive count for AT LEAST ONE of 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

GQ Type-Level

GQ ID	GQ Type	Mean # of GQs	Ratio Adjusted		Poisson		Median		IPEDS	
			Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
US	National	9,200	140,000	141,100			-48,500	13,240		
104	Local Jails and Other Municipal Confinement Facilities	70	-90	2,890			-300	2,685		
105	Correctional Residential Facilities	40	450	1,313			-100	1,136		
106	Military Disciplinary Barracks and Jails	N<15	20	56			-20	272		
201	Group Homes for Juveniles (non-correctional)	150	11,000	34,800			-200	177		
202	Residential Treatment Centers for Juveniles (non-correctional)	80	90	196			-60	243		
203	Correctional Facilities Intended for Juveniles (training schoo	40	150	323			-150	287		
301	Nursing Facilities/Skilled-Nursing Facilities	1000	1,500	4,002			-6,200	2,771		
301	Nursing Facilities/Skilled-Nursing Facilities (Total Number of Occupied Beds)	100	-400	301			-4,600	936	-700	506
301	Nursing Facilities/Skilled-Nursing Facilities (Number of All Beds)	100	-400	298			-4,600	931	2,600	462
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	40	200	580			-100	922		
402	Hospitals with Patients Who Have No Usual Home Elsewhere	20	100	226			50	187		

403	In-Patient Hospice Facilities	30	-N<15	N<15			-90	232		
404	Military Treatment Facilities with Assigned Active Duty Patien	N<15	-50	75			60	183		
405	Residential Schools for People with Disabilities	20	150	238			20	183		
501	College/University Student Housing	1,400	28,000	54,971			-26,000	9,530		
501	College/University Student Housing – universe reduced for IPEDS	1,400	28,000	54,968			-26,000	9,551	86,500	62,320
601	Military Quarters	150	24,000	41,715			2,000	6,400		
602	Military Ships	30	50	1,318			-1,300	2,800		
701	Emergency and transitional shelters (with sleeping facilities)	300	300	901			-1,500	1,127		
702	Soup Kitchens	100	650	2,898			-550	1,131		
704	Regularly Scheduled Mobile Food Vans	20	30	238			-40	438		
706	Targeted Non-Sheltered Outdoor Locations	700	1,900	1,121			-2,200	1,004		
801	Group Homes Intended for Adults (non-correctional)	3,000	24,500	39,792			-10,500	1,189		
802	Residential Treatment Centers for Adults (non-correctional)	450	300	723			-1,500	704		
900	Maritime/Merchant Vessels	40	N<15	N<15			-50	93		
901	Workers' Group Living Quarters and Job Corps Centers	350	400	2,130			-2,200	1,554		
903	Living Quarters for Victims of Natural Disasters	N<15	30	103			N<15	N<15		
999	Other	50	100	323			-400	273		

<u>Test 3: We nothing from 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size</u>

<u>GQ Type-Level</u>

GQ ID	GQ Type	Mean # of GQs	Ratio Adjusted		Poisson		Median		IPEDS	
			Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
US	National	2,200					1,600	3,889		
104	Local Jails and Other Municipal Confinement Facilities	N<15					N<15	N<15		
105	Correctional Residential Facilities	20					-300	522		
106	Military Disciplinary Barracks and Jails									
201	Group Homes for Juveniles (non-correctional)	N<15					-20	39		
202	Residential Treatment Centers for Juveniles (non-correctional)	N<15					N<15	N<150		
203	Correctional Facilities Intended for Juveniles (training schoo	N<15					30	267		
301	Nursing Facilities/Skilled-Nursing Facilities	40					-250	222		
301	Nursing Facilities/Skilled-Nursing Facilities (Total Number of Occupied Beds)	N<15					-100	74	-30	68

301	Nursing Facilities/Skilled-Nursing Facilities (Number of All Beds)	N<15					-100	74	N<15	N<15
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	N<15					N<15	N<15		
402	Hospitals with Patients Who Have No Usual Home Elsewhere	N<15					N<15	N<15		
403	In-Patient Hospice Facilities	N<15					-20	62		
404	Military Treatment Facilities with Assigned Active Duty Patien									
405	Residential Schools for People with Disabilities	N<15					N<15	N<15		
501	College/University Student Housing	100					-250	2,938		
501	College/University Student Housing – universe reduced for IPEDS	100					-450	2,911	15,500	3,745
601	Military Quarters	40					-200	509		
602	Military Ships									
701	Emergency and transitional shelters (with sleeping facilities)	40					-20	283		
702	Soup Kitchens	20					-20	763		
704	Regularly Scheduled Mobile Food Vans	N<15					N<15	N<15		
706	Targeted Non-Sheltered Outdoor Locations	950					-2,100	519		
801	Group Homes Intended for Adults (non-correctional)	200					-350	363		
802	Residential Treatment Centers for Adults (non-correctional)	30					50	176		
900	Maritime/Merchant Vessels									
901	Workers' Group Living Quarters and Job Corps Centers	20					-70	158		
903	Living Quarters for Victims of Natural Disasters	N<15					-30	.		
999	Other	20					-150	358		

Andrew Keller
Imputing GQ Pop Counts
December 20, 2020

Input Data – Does not incorporate Call-in or web scraping results

Table 1: Input Data

GQ Status	No Good Person (GP)	Has Good Person	Total
Occupied GQ	17,000	181,000	197,000
Delete GQ	7,200	450	7,600
Nonresidential GQ	2,400	100	2,500
Vacant During Visit, Open on Census Day	19,500	1,900	21,500
Refusal GQ	6,700	1,100	7,800
Vacant GQ	29,000	1,100	30,500
Total	82,000	185,000	267,000

Make 10 replicates each of 10% missingness – using Juli's indicators for cases that should be suppressed or imputed.

<u>Test 1: We have a positive count for ALL 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size</u>

Four Methods:

1. Ratio-Adjusted 2020 GQ Advanced Contact (AC) Expected Count – Needs GQ AC count
   a. If we have 2020 GQAC Expected count, we adjust it by a ratio determined by Good Person / GQAC ratio within state and GQ Type

2. Poisson Model –
   a. Model Count offset GQ Current Max Size

```
proc genmod data = nomaxmod;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

3. Take variable percentile good person count with state and GQ type as imputed count – Needs nothing
   a. 104, 801, 802, 901 – use 70th
   b. 301 – use 55th
   c. 501 – use 68th
   d. All others – use 65th
4. CES IPEDS method

<u>Test 2: We have a positive count for AT LEAST ONE of 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size</u>

Three Methods:

1. Ratio-adjustment with hierarchy:
   a. Ratio-Adjusted 2020 GQ Advanced Contact (AC) Expected Count – Needs GQ AC count

   b. Ratio-Adjusted 2020 GQAC Max Size – Needs GQ AC Max Size
   c. Ratio-Adjusted Current Surveys Count – Needs Current Survey count
   d. Ratio-Adjusted Current Surveys Max Size – Needs Current Survey Max Size

2. Take variable percentile good person count with state and GQ type as imputed count – Needs nothing
3. CES IPEDS method

Test 3: We nothing from 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

Two Methods:

1. Take variable percentile good person count with state and GQ type as imputed count – Needs nothing
2. CES IPEDS method

Results

Applying Models to 12/17 Truth Decks
Bias = SUM(Imputed GQ Pop) – SUM(Provided GQ Pop)

Test 1: We have a positive count for ALL 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

GQ Type-Level

GQ ID	GQ Type PUT IN MEDIAN BIAS	Mean # of GQs	Ratio Adjusted		Poisson		Median		IPEDS	
			Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
US	National	6,500	3,700	3,110	3,400	3,356	-6,900	11,920		
104	Local Jails and Other Municipal Confinement Facilities	200	400	3,152	100	2,724	-1,200	3,456		
105	Correctional Residential Facilities	30	20	396	20	485	-150	671		
106	Military Disciplinary Barracks and Jails	N<15	N<15	N<15	N<15	N<15	N<15	N<15		
201	Group Homes for Juveniles (non-correctional)	150	90	96	80	135	-300	157		
202	Residential Treatment Centers for Juveniles (non-correctional)	90	60	137	40	153	-150	191		
203	Correctional Facilities Intended for Juveniles (training schoo	60	100	426	30	233	-50	189		
301	Nursing Facilities/Skilled-Nursing Facilities	1,400	750	1,300	750	1,715	-1,200	2,265		
301	Nursing Facilities/Skilled-Nursing Facilities (Total Number of Occupied Beds)	700	150	694	300	1,025	-6,500	1,681	-3,600	585
301	Nursing Facilities/Skilled-Nursing Facilities (Number of All Beds)	700	150	694	300	1,025	-6,500	1,681	16,500	1,201
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	30	90	435	N<15	N<15	-200	1,081		
402	Hospitals with Patients Who Have No Usual Home Elsewhere	N<15	30	62	N<15	N<15	40	171		
403	In-Patient Hospice Facilities	20	30	50	N<15	N<15	-20	102		
404	Military Treatment Facilities with Assigned Active Duty Patien	N<15	N<15	N<15	N<15	N<15	N<15	N<15		

405	Residential Schools for People with Disabilities	20	70	104	N<15	N<15	-20	400		
501	College/University Student Housing	1,300	600	2,244	700	2,520	2,100	9,474		
501	College/University Student Housing – universe reduced for IPEDS	1,300	600	2,244	700	2,520	2,100	9,474	8,200	1,726
601	Military Quarters	80	-60	1,315	N<15	N<15	-60	2,229		
701	Emergency and transitional shelters (with sleeping facilities)	150	150	429	80	407	-750	733		
702	Soup Kitchens	70	80	605	150	588	-100	948		
704	Regularly Scheduled Mobile Food Vans	N<15	N<15	N<15	N<15	N<15	-20	36		
706	Targeted Non-Sheltered Outdoor Locations	20	20	102	N<15	N<15	-30	119		
801	Group Homes Intended for Adults (non-correctional)	2,100	850	319	1,100	658	-2,600	487		
802	Residential Treatment Centers for Adults (non-correctional)	350	200	217	200	199	-800	463		
901	Workers' Group Living Quarters and Job Corps Centers	250	150	250	100	248	-750	739		
903	Living Quarters for Victims of Natural Disasters	N<15	N<15	N<15	N<15	N<15	20	106		
999	Other	N<15	N<15	53	N<15	N<15	-20	39		

Test 2: We have a positive count for AT LEAST ONE of 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

GQ Type-Level

GQ ID	GQ Type	Mean # of GQs	Ratio Adjusted		Poisson		Median		IPEDS	
			Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
US	National	9,200	140,000	141,100			-48,500	13,240		
104	Local Jails and Other Municipal Confinement Facilities	70	-90	2,890			-300	2,685		
105	Correctional Residential Facilities	40	450	1,313			-100	1,136		
106	Military Disciplinary Barracks and Jails	N<15	20	56			-20	272		
201	Group Homes for Juveniles (non-correctional)	150	11,000	34,800			-200	177		
202	Residential Treatment Centers for Juveniles (non-correctional)	80	90	196			-60	243		
203	Correctional Facilities Intended for Juveniles (training schoo	40	150	323			-150	287		
301	Nursing Facilities/Skilled-Nursing Facilities	1000	1,500	4,002			-6,200	2,771		
301	Nursing Facilities/Skilled-Nursing Facilities (Total Number of Occupied Beds)	100	-400	301			-4,600	936	-700	506
301	Nursing Facilities/Skilled-Nursing Facilities (Number of All Beds)	100	-400	298			-4,600	931	2,600	462
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	40	200	580			-100	922		
402	Hospitals with Patients Who Have No Usual Home Elsewhere	20	100	226			50	187		

403	In-Patient Hospice Facilities	30	-N<15	N<15			-90	232		
404	Military Treatment Facilities with Assigned Active Duty Patien	N<15	-50	75			60	183		
405	Residential Schools for People with Disabilities	20	150	238			20	183		
501	College/University Student Housing	1,400	28,000	54,971			-26,000	9,530		
501	College/University Student Housing – universe reduced for IPEDS	1,400	28,000	54,968			-26,000	9,551	86,500	62,320
601	Military Quarters	150	24,000	41,715			2,000	6,400		
602	Military Ships	30	50	1,318			-1,300	2,800		
701	Emergency and transitional shelters (with sleeping facilities)	300	300	901			-1,500	1,127		
702	Soup Kitchens	100	650	2,898			-550	1,131		
704	Regularly Scheduled Mobile Food Vans	20	30	238			-40	438		
706	Targeted Non-Sheltered Outdoor Locations	700	1,900	1,121			-2,200	1,004		
801	Group Homes Intended for Adults (non-correctional)	3,000	24,500	39,792			-10,500	1,189		
802	Residential Treatment Centers for Adults (non-correctional)	450	300	723			-1,500	704		
900	Maritime/Merchant Vessels	40	N<15	N<15			-50	93		
901	Workers' Group Living Quarters and Job Corps Centers	350	400	2,130			-2,200	1,554		
903	Living Quarters for Victims of Natural Disasters	N<15	30	103			N<15	N<15		
999	Other	50	100	323			-400	273		

Test 3: We nothing from 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

GQ Type-Level

GQ ID	GQ Type	Mean # of GQs	Ratio Adjusted		Poisson		Median		IPEDS	
			Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
US	National	2,200					1,600	3,889		
104	Local Jails and Other Municipal Confinement Facilities	N<15					N<15	N<15		
105	Correctional Residential Facilities	20					-300	522		
106	Military Disciplinary Barracks and Jails									
201	Group Homes for Juveniles (non-correctional)	N<15					-20	39		
202	Residential Treatment Centers for Juveniles (non-correctional)	N<15					N<15	N<15		
203	Correctional Facilities Intended for Juveniles (training schoo	N<15					30	267		
301	Nursing Facilities/Skilled-Nursing Facilities	40					-250	222		
301	Nursing Facilities/Skilled-Nursing Facilities (Total Number of Occupied Beds)	N<15					-100	74	-30	68

301	Nursing Facilities/Skilled-Nursing Facilities (Number of All Beds)	N<15					-100	74	N<15	N<15
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	N<15					N<15	N<15		
402	Hospitals with Patients Who Have No Usual Home Elsewhere	N<15					N<15	N<15		
403	In-Patient Hospice Facilities	N<15					-20	62		
404	Military Treatment Facilities with Assigned Active Duty Patien									
405	Residential Schools for People with Disabilities	N<15					N<15	N<15		
501	College/University Student Housing	100					-250	2,938		
501	College/University Student Housing – universe reduced for IPEDS	100					-450	2,911	15,500	3,745
601	Military Quarters	40					-200	509		
602	Military Ships									
701	Emergency and transitional shelters (with sleeping facilities)	40					-20	283		
702	Soup Kitchens	20					-20	763		
704	Regularly Scheduled Mobile Food Vans	N<15					N<15	N<15		
706	Targeted Non-Sheltered Outdoor Locations	950					-2,100	519		
801	Group Homes Intended for Adults (non-correctional)	200					-350	363		
802	Residential Treatment Centers for Adults (non-correctional)	30					50	176		
900	Maritime/Merchant Vessels									
901	Workers' Group Living Quarters and Job Corps Centers	20					-70	158		
903	Living Quarters for Victims of Natural Disasters	N<15					-30	.		
999	Other	20					-150	358		

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards

Andrew Keller
Imputing GQ Pop Counts
December 20, 2020

Input Data – Does not incorporate Call-in or web scraping results

Table 1: Input Data

GQ Status	No Good Person (GP)	Has Good Person	Total
Occupied GQ	17,000	181,000	197,000
Delete GQ	7,200	450	7,600
Nonresidential GQ	2,400	100	2,500
Vacant During Visit, Open on Census Day	19,500	1,900	21,500
Refusal GQ	6,700	1,100	7,800
Vacant GQ	29,000	1,100	30,500
Total	82,000	185,000	267,000

Make 10 replicates each of 10% missingness – using Juli's indicators for cases that should be suppressed or imputed.

Test 1: We have a positive count for ALL 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

Four Methods:

1. Ratio-Adjusted 2020 GQ Advanced Contact (AC) Expected Count – Needs GQ AC count
    a. If we have 2020 GQAC Expected count, we adjust it by a ratio determined by Good Person / GQAC ratio within state and GQ Type

2. Poisson Model –
    a. Model Count offset GQ Current Max Size

```
proc genmod data = nomaxmod;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

3. Take variable percentile good person count with state and GQ type as imputed count – Needs nothing
    a. 104, 801, 802, 901 – use 70th
    b. 301 – use 55th
    c. 501 – use 68th
    d. All others – use 65th
4. CES IPEDS method

Test 2: We have a positive count for AT LEAST ONE of 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

Three Methods:

1. Ratio-adjustment with hierarchy:
    a. Ratio-Adjusted 2020 GQ Advanced Contact (AC) Expected Count – Needs GQ AC count

   b. Ratio-Adjusted 2020 GQAC Max Size – Needs GQ AC Max Size
   c. Ratio-Adjusted Current Surveys Count – Needs Current Survey count
   d. Ratio-Adjusted Current Surveys Max Size – Needs Current Survey Max Size

2. Take variable percentile good person count with state and GQ type as imputed count – Needs nothing
3. CES IPEDS method

Test 3: We nothing from 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

Two Methods:

1. Take variable percentile good person count with state and GQ type as imputed count – Needs nothing
2. CES IPEDS method

Results

Applying Models to 12/17 Truth Decks
Bias = SUM(Imputed GQ Pop) – SUM(Provided GQ Pop)

Test 1: We have a positive count for ALL 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size **(Note that ratio-adjusted by 2020 GQAC Max Count and 2020 Current Surveys Max Count are the same. This is because they have the same value if both are filled.)**

GQ Type-Level

GQ ID	GQ Type	Mean # of GQs	Ratio Adjusted By 2020 GQAC Expected Count		Ratio Adjusted By 2020 GQAC Max Count		Ratio Adjusted By 2020 Current Surveys Count		Ratio Adjusted By 2020 Current Surveys Max Count		Poisson		Median		CES	
			Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
US	National	6,500	3,700	3,110							3,400	3,356	-6,900	11,920		
104	Local Jails and Other Municipal Confinement Facilities	200	400	3,152	750	4,475	300	2,106	750	4,475	100	2,724	-1,200	3,456		
105	Correctional Residential Facilities	30	20	396	-100	580	600	2,018	-100	580	20	485	-150	671		
106	Military Disciplinary Barracks and Jails	N<15	N<15	N<15	N<15	71	100	89	N<15	N<15	N<15	N<15	N<15	N<15		
201	Group Homes for Juveniles (non-correctional)	150	90	96	100	158	90	163	100	158	80	135	-300	157		
202	Residential Treatment Centers for Juveniles (non-correctional)	90	60	137	650	1,556	80	211	650	1,556	40	153	-150	191		
203	Correctional Facilities Intended for Juveniles (training schoo	60	100	426	100	327	150	305	100	327	30	233	-50	189		
301	Nursing Facilities/Skilled-Nursing Facilities	1,400	750	1,300	1,400	8,941	750	1,196	1,400	8,941	750	1,715	-1,200	2,265		
301	Nursing Facilities/Skilled-Nursing Facilities (Total_Number_of_Occupied Beds)	700	150	694							300	1,025	-6,500	1,681	-3,600	585
301	Nursing Facilities/Skilled-Nursing Facilities (Number of All Beds)	700	150	694							300	1,025	-6,500	1,681	16,500	1,201

401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	30	90	435	50	420	650	2,275	50	420	N<15	N<15	-200	1,081		
402	Hospitals with Patients Who Have No Usual Home Elsewhere	N<15	30	62	30	80	60	133	30	80	N<15	N<15	40	171		
403	In-Patient Hospice Facilities	20	30	50	20	40	40	90	20	40	N<15	N<15	-20	102		
404	Military Treatment Facilities with Assigned Active Duty Patien	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15		
405	Residential Schools for People with Disabilities	20	70	104	80	87	100	444	80	87	N<15	N<15	-20	400		
501	College/University Student Housing	1,300	600	2,244	2,800	11,670	700	2,538	2,800	11,670	700	2,520	2,100	9,474		
501	College/University Student Housing – universe reduced for IPEDS	1,300	600	2,244							700	2,520	2,100	9,474	8,200	1,726
601	Military Quarters	80	-60	1,315	-40	1,522	250	1,714	-40	1,522	N<15	N<15	-60	2,229		
701	Emergency and transitional shelters (with sleeping facilities)	150	150	429	150	404	150	495	150	404	80	407	-750	733		
702	Soup Kitchens	70	80	605	200	1,180	550	2,011	200	1,180	150	588	-100	948		
704	Regularly Scheduled Mobile Food Vans	N<15	N<15	19	N<15	N<15	N<15	27	N<15	43	N<15	N<15	-20	36		
706	Targeted Non-Sheltered Outdoor Locations	20	20	102	20	103	70	266	20	103	N<15	N<15	-30	119		
801	Group Homes Intended for Adults (non-correctional)	2,100	850	319	1,100	809	1,100	457	1,100	809	1,100	658	-2,600	487		
802	Residential Treatment Centers for Adults (non-correctional)	350	200	217	200	217	250	389	200	217	200	199	-800	463		
901	Workers' Group Living Quarters and Job Corps Centers	250	150	250	150	320	300	671	150	320	100	248	-750	739		
903	Living Quarters for Victims of Natural Disasters	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	20	106		
999	Other	N<15	20	53	20	46	N<15	N<15	20	46	N<15	N<15	-20	39		

Test 2: We have a positive count for AT LEAST ONE of 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

GQ Type-Level

GQ ID	GQ Type	Mean # of GQs	Ratio Adjusted		Poisson		Median		IPEDS	
			Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
US	National	9,200	140,000	141,100			-48,500	13,240		
104	Local Jails and Other Municipal Confinement Facilities	70	-90	2,890			-300	2,685		
105	Correctional Residential Facilities	40	450	1,313			-100	1,136		
106	Military Disciplinary Barracks and Jails	N<15	20	56			-20	272		

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

201	Group Homes for Juveniles (non-correctional)	150	11,000	34,800			-200	177		
202	Residential Treatment Centers for Juveniles (non-correctional)	80	90	196			-60	243		
203	Correctional Facilities Intended for Juveniles (training schoo	40	150	323			-150	287		
301	Nursing Facilities/Skilled-Nursing Facilities	1000	1,500	4,002			-6,200	2,771		
301	Nursing Facilities/Skilled-Nursing Facilities (Total Number of Occupied Beds)	100	-400	301			-4,600	936	-700	506
301	Nursing Facilities/Skilled-Nursing Facilities (Number of All Beds)	100	-400	298			-4,600	931	2,600	462
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	40	200	580			-100	922		
402	Hospitals with Patients Who Have No Usual Home Elsewhere	20	100	226			50	187		
403	In-Patient Hospice Facilities	30	N<15	N<15			-90	232		
404	Military Treatment Facilities with Assigned Active Duty Patien	N<15	-50	75			60	183		
405	Residential Schools for People with Disabilities	20	150	238			20	183		
501	College/University Student Housing	1,400	28,000	54,970			-26,000	9,530		
501	College/University Student Housing – universe reduced for IPEDS	1,400	28,000	54,970			-26,000	9,551	86,500	62,320
601	Military Quarters	150	24,000	41,720			2,000	6,400		
602	Military Ships	30	50	1,318			-1,300	2,800		
701	Emergency and transitional shelters (with sleeping facilities)	300	300	901			-1,500	1,127		
702	Soup Kitchens	100	650	2,898			-550	1,131		
704	Regularly Scheduled Mobile Food Vans	20	30	238			-40	438		
706	Targeted Non-Sheltered Outdoor Locations	700	1,900	1,121			-2,200	1,004		
801	Group Homes Intended for Adults (non-correctional)	3,000	24,500	39,790			-10,500	1,189		
802	Residential Treatment Centers for Adults (non-correctional)	450	300	723			-1,500	704		
900	Maritime/Merchant Vessels	40	N<15	N<15			-50	93		
901	Workers' Group Living Quarters and Job Corps Centers	350	400	2,130			-2,200	1,554		
903	Living Quarters for Victims of Natural Disasters	N<15	30	103			N<15	N<15		
999	Other	50	100	323			-400	273		

Test 3: We nothing from 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

GQ Type-Level

GQ ID	GQ Type	Mean # of GQs	Ratio Adjusted		Poisson		Median		IPEDS	
			Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
US	National	2,200					1,600	3,889		

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

104	Local Jails and Other Municipal Confinement Facilities	N<15					N<15	N<15		
105	Correctional Residential Facilities	20					-300	522		
106	Military Disciplinary Barracks and Jails									
201	Group Homes for Juveniles (non-correctional)	N<15					-20	39		
202	Residential Treatment Centers for Juveniles (non-correctional)	N<15					N<15	N<15		
203	Correctional Facilities Intended for Juveniles (training schoo	N<15					30	267		
301	Nursing Facilities/Skilled-Nursing Facilities	40					-250	222		
301	Nursing Facilities/Skilled-Nursing Facilities (Total Number of Occupied Beds)	N<15					-100	74	-30	68
301	Nursing Facilities/Skilled-Nursing Facilities (Number of All Beds)	N<15					-100	74	N<15	N<15
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	N<15					N<15	N<15		
402	Hospitals with Patients Who Have No Usual Home Elsewhere	N<15					N<15	N<15		
403	In-Patient Hospice Facilities	N<15					-20	62		
404	Military Treatment Facilities with Assigned Active Duty Patien									
405	Residential Schools for People with Disabilities	N<15					N<15	N<15		
501	College/University Student Housing	100					-250	2,938		
501	College/University Student Housing – universe reduced for IPEDS	100					-450	2,911	15,500	3,745
601	Military Quarters	40					-200	509		
602	Military Ships									
701	Emergency and transitional shelters (with sleeping facilities)	40					-20	283		
702	Soup Kitchens	20					-20	763		
704	Regularly Scheduled Mobile Food Vans	N<15					N<15	N<15		
706	Targeted Non-Sheltered Outdoor Locations	950					-2,100	519		
801	Group Homes Intended for Adults (non-correctional)	200					-350	363		
802	Residential Treatment Centers for Adults (non-correctional)	30					50	176		
900	Maritime/Merchant Vessels									
901	Workers' Group Living Quarters and Job Corps Centers	20					-70	158		
903	Living Quarters for Victims of Natural Disasters	N<15					-30	.		
999	Other	20					-150	358		

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Andrew Keller
Imputing GQ Pop Counts
December 20, 2020

Input Data – Does not incorporate Call-in or web scraping results

Table 1: Input Data

GQ Status	No Good Person (GP)	Has Good Person	Total
Occupied GQ	17,000	181,000	197,000
Delete GQ	7,200	450	7,600
Nonresidential GQ	2,400	100	2,500
Vacant During Visit, Open on Census Day	19,500	1,900	21,500
Refusal GQ	6,700	1,100	7,800
Vacant GQ	29,000	1,100	30,500
Total	82,000	185,000	267,000

Make 10 replicates each of 10% missingness – using Juli's indicators for cases that should be suppressed or imputed.

Test 1: We have a positive count for ALL 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

Four Methods:

1. Ratio-Adjusted 2020 GQ Advanced Contact (AC) Expected Count – Needs GQ AC count
   a. If we have 2020 GQAC Expected count, we adjust it by a ratio determined by Good Person / GQAC ratio within state and GQ Type

2. Poisson Model –
   a. Model Count offset GQ Current Max Size

```
proc genmod data = nomaxmod;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

3. Take variable percentile good person count with state and GQ type as imputed count – Needs nothing
   a. 104, 801, 802, 901 – use 70th
   b. 301 – use 55th
   c. 501 – use 68th
   d. All others – use 65th
4. CES IPEDS method

Test 2: We have a positive count for AT LEAST ONE of 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

Three Methods:

1. Ratio-adjustment with hierarchy:
   a. Ratio-Adjusted 2020 GQ Advanced Contact (AC) Expected Count – Needs GQ AC count

b. Ratio-Adjusted 2020 GQAC Max Size – Needs GQ AC Max Size
c. Ratio-Adjusted Current Surveys Count – Needs Current Survey count
d. Ratio-Adjusted Current Surveys Max Size – Needs Current Survey Max Size

2. Take variable percentile good person count with state and GQ type as imputed count – Needs nothing
3. CES IPEDS method

Test 3: We nothing from 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

Two Methods:

1. Take variable percentile good person count with state and GQ type as imputed count – Needs nothing
2. CES IPEDS method

Results

Applying Models to 12/17 Truth Decks
Bias = SUM(Imputed GQ Pop) – SUM(Provided GQ Pop)

Test 1: We have a positive count for ALL 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size **(Note that ratio-adjusted by 2020 GQAC Max Count and 2020 Current Surveys Max Count are the same. This is because they have the same value if both are filled.)**

GQ Type-Level

GQ ID	GQ Type	Mean # of GQs	Ratio Adjusted By 2020 GQAC Expected Count		Ratio Adjusted By 2020 GQAC Max Count		Ratio Adjusted By 2020 Current Surveys Count		Ratio Adjusted By 2020 Current Surveys Max Count		Poisson		Median		CES	
			Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
US	National															
104	Local Jails and Other Municipal Confinement Facilities	200	150	1,950	300	2,100	250	1,334	700	2,166	200	1,690	-1,200	2,490		
105	Correctional Residential Facilities	30	100	463	70	318	-70	280	100	374	30	321	-70	706		
106	Military Disciplinary Barracks and Jails	N<15	N<15	N<15	N<15	N<15	-30	62	-20	51	N<15	N<15	30	146		
201	Group Homes for Juveniles (non-correctional)	150	50	132	70	139	80	123	100	145	70	111	-200	141		
202	Residential Treatment Centers for Juveniles (non-correctional)	90	30	105	40	141	30	137	20	133	50	99	-90	271		
203	Correctional Facilities Intended for Juveniles (training schoo	60	90	224	200	294	150	239	200	277	30	168	N<15	N<158		
301	Nursing Facilities/Skilled-Nursing Facilities	1,400	300	981	150	623	700	698	-60	650	750	844	-750	1,547		
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	30	150	342	100	357	N<15	N<15	N<15	N<15	20	217	-80	463		
402	Hospitals with Patients Who Have No Usual Home Elsewhere	N<15	30	61	20	59	50	65	30	105	N<15	N<15	40	209		
403	In-Patient Hospice Facilities	20	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15		

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

404	Military Treatment Facilities with Assigned Active Duty Patien	N<15	N<15	N<15	N<15	N<15	20	3	N<15	N<15	N<15	N<15	N<15	N<15		
405	Residential Schools for People with Disabilities	20	70	179	90	206	N<15	N<15	80	205	20	154	N<15	N<15		
501	College/University Student Housing	1,300	-20	1,183	1,400	1,853	600	1,462	2,000	1,839	650	1,145	2,800	7,957	-350	2,388
601	Military Quarters	80	300	1,493	350	1,113	350	712	150	1,019	60	1,051	300	1,192		
701	Emergency and transitional shelters (with sleeping facilities)	150	150	320	200	333	N<15	N<15	450	333	80	388	-750	769		
702	Soup Kitchens	70	250	453	550	585	150	827	950	563	40	642	-250	691		
704	Regularly Scheduled Mobile Food Vans	N<15	N<15	N<15	80	240	N<15	N<15	90	242	N<15	N<15	-20	63		
706	Targeted Non-Sheltered Outdoor Locations	20	20	100	100	123	70	171	20	97	30	153	-20	64		
801	Group Homes Intended for Adults (non-correctional)	2,100	400	304	300	302	900	317	350	305	1,100	310	-1,900	663		
802	Residential Treatment Centers for Adults (non-correctional)	350	200	281	200	313	100	280	300	334	200	336	-650	676		
901	Workers' Group Living Quarters and Job Corps Centers	200	200	237	150	214	100	327	200	166	100	175	-600	549		
903	Living Quarters for Victims of Natural Disasters	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	-20	36		
999	Other	N<15	30	66	N<15	N<15	N<15	N<15	N<15	N<15	N<15	N<15	-20	42		

Test 2: We have a positive count for AT LEAST ONE of 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

GQ Type-Level

GQ ID	GQ Type	Mean # of GQs	Ratio Adjusted		Poisson		Median		IPEDS	
			Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
US	National									
104	Local Jails and Other Municipal Confinement Facilities	40	-50	836			1,100	2,783		
105	Correctional Residential Facilities	N<15	50	113			-300	1,126		
106	Military Disciplinary Barracks and Jails	150	70	105			30	173		
201	Group Homes for Juveniles (non-correctional)	70	70	76			70	108		
202	Residential Treatment Centers for Juveniles (non-correctional)	40	-70	181			350	198		
203	Correctional Facilities Intended for Juveniles (training schoo	950	850	780			-100	254		
301	Nursing Facilities/Skilled-Nursing Facilities	70	-550	1,202			24,500	3,372		
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	40	250	363			1,600	664		

402	Hospitals with Patients Who Have No Usual Home Elsewhere	N<15	100	181			90	155		
403	In-Patient Hospice Facilities	20	N<15	N<15			-30	237		
404	Military Treatment Facilities with Assigned Active Duty Patien	N<15	-30	99			-60	85		
405	Residential Schools for People with Disabilities	20	N<15	N<15			250	118		
501	College/University Student Housing	1,400	-1,200	2,892			24,000	12,350	~~1,000~~ 1,000	~~4,600~~ 4,600
601	Military Quarters	150	650	3,392			3,200	4,255		
602	Military Ships	20	-200	1,064			-2,400	3,437		
701	Emergency and transitional shelters (with sleeping facilities)	300	-400	778			-2,000	753		
702	Soup Kitchens	90	-1000	460			90	524		
704	Regularly Scheduled Mobile Food Vans	20	-50	259			-350	374		
706	Targeted Non-Sheltered Outdoor Locations	700	1,400	1,032			-1,200	964		
801	Group Homes Intended for Adults (non-correctional)	3,000	1,800	922			-7,200	1,183		
802	Residential Treatment Centers for Adults (non-correctional)	450	-20	406			-200	935		
900	Maritime/Merchant Vessels	40	20	74			1,200	687		
901	Workers' Group Living Quarters and Job Corps Centers	300	150	361			-450	1,364		
903	Living Quarters for Victims of Natural Disasters	N<15	N<15	N<15			N<15	19		
999	Other	50	90	159			-600	273		

Test 3: We nothing from 2020 GQAC Expected count, 2020 GQAC Max Size, Current Surveys Count, Current Surveys Max Size

GQ Type-Level

GQ ID	GQ Type	Mean # of GQs	Ratio Adjusted		Poisson		Median		IPEDS	
			Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)	Mean(Bias)	SE(Bias)
US	National									
104	Local Jails and Other Municipal Confinement Facilities	N<15					350	304		
105	Correctional Residential Facilities	20					300	932		
106	Military Disciplinary Barracks and Jails									
201	Group Homes for Juveniles (non-correctional)	N<15					(D)	(D)		
202	Residential Treatment Centers for Juveniles (non-correctional)	N<15					80	204		
203	Correctional Facilities Intended for Juveniles (training schoo	N<15					60	167		
301	Nursing Facilities/Skilled-Nursing Facilities	40					1,800	893		

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other	N<15					200	114		
402	Hospitals with Patients Who Have No Usual Home Elsewhere	N<15					50	122		
403	In-Patient Hospice Facilities	N<15					-20	63		
404	Military Treatment Facilities with Assigned Active Duty Patien									
405	Residential Schools for People with Disabilities	N<15					30	53		
501	College/University Student Housing	100					2,400	3,118	~~1,000~~ -30	~~2,484~~ 4,007
601	Military Quarters	40					5,600	5,975		
602	Military Ships	40					50	357		
701	Emergency and transitional shelters (with sleeping facilities)	N<15					30	429		
702	Soup Kitchens	20					N<15	N<15		
704	Regularly Scheduled Mobile Food Vans	N<15					1,800	1,091		
706	Targeted Non-Sheltered Outdoor Locations	950					40	296		
801	Group Homes Intended for Adults (non-correctional)	200					150	168		
802	Residential Treatment Centers for Adults (non-correctional)	30					30	53		
900	Maritime/Merchant Vessels									
901	Workers' Group Living Quarters and Job Corps Centers	20					-30	186		
903	Living Quarters for Victims of Natural Disasters	N<15					-30			
999	Other	20					-300	370		

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Variables included in the Group Quarters File (FULL_DATASET4) provided by DSSD

Ryan King, Maranda Pepe, Hannah Zimmerman, Mary Frances Zelenak

December 8, 2020

Variable	Definition	Source	Variable Type and Length
ACOCE	Area Census Office	Universe File	Char $16.
BCUCOUNTYFP	FIPS County Code	Universe File	Char $12.
BCUSTATEFP	FIPS State Code	Universe File	Char $8.
CDLPER	Indicator for Persons found in CDL, on 10/27.	CDL	Num 8
CDLPER_LATE	Indicator for Persons Found in CDL run on 12/8/20.	CDL	Num 8
DDP	Indicator for data-defined person.	DRF1	Num 8
DDP_PSA	Approximate number of data-defined people after preliminary DRF2 processing to resolve cases with more than one response at a MAFID.	** Post-DRF2 Processing Test File	Num 8
DRPS_CDL	Person records sent to both DRPS and CDL.	DRF1, CDL	Char 1
ER_STATED_CNT	The number of people residing in the Group Quarters on Census day, as stated by the Group Quarters contact person (administrator). Recorded during eResponse.	Case-Level File	Num 8 6.
FACTLNAME	The name of the Group Quarters facility. A facility is an umbrella organization that has a group of GQs. For example, a college is the facility and its dorms are the GQs.	Universe File	Char $400.
FOCS_CALCULATED	The number of people residing in the Group Quarters on Census day determined by the number of person responses recorded in FOCS.	Case-Level File	Num 8 6.
FOCS_ER_CB_CODE	A code that identifies the FOCS disposition status of the enumeration record associated with the Group Quarters.	Case-Level File	Char $12.
FOCS_STATED_CNT	The number of people residing in the Group Quarters on Census day, as stated by the Group Quarters contact person (administrator). Recorded in FOCS.	Case-Level File	Num 8 6.
GP	Number of people with good person flag, GP = 1	DRF1	Num 8
GP_PSA	Approximate number of good people (GP=1) after preliminary DRF2 processing to resolve cases with more than one response at a MAFID.	** Post-DRF2 Processing Test File	Num 8
GP13	Number of CDL people from Session_context_code 13, eResponse.	CDL	Num 8
GP15	Number of CDL people from Session_context_code 15, Individual Census Questionnaires	CDL	Num 8
GP17	Number of CDL peoplefrom Session_context_code 17, Paper Listings.	CDL	Num 8

GP13_LATE	all versions of the previous CDL indicators	CDL	Num 8
GP15_LATE	all versions of the previous CDL indicators	CDL	Num 8
UGP15_LATE	Number of unlinked ICQs from CDL, there is overlap with GP15_LATE.	CDL	Num 8
GP17_LATE	all versions of the previous CDL indicators	CDL	Num 8
GQ_OC_CNT_0917	Counts from GQ Off-Campus work on 9/17	Off-Campus Records	Num 8
GQ_OC_CNT_0924	Counts from GQ Off-Campus work on 9/24	Off-Campus Records	Num 8
GQ_PRI_PH_AREA_ID	The unique identifier of the phone number area code of the Group Quarters primary point of contact.	GQ Advanced Contact (GQAC)	Char $3.
GQ_PRI_PH_EXT_TEXT	The phone number extension of the Group Quarters primary point of contact.	GQ Advanced Contact (GQAC)	Char $8
GQ_PRI_PH_LINE_TEXT	The phone number line of the Group Quarters primary point of contact.	GQ Advanced Contact (GQAC)	Char $7.
GQ_SIZE_EXP_PERS_CNT	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact (GQAC)	Num 8 6.
GQ_SIZE_MAX_PERS_CNT	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact (GQAC)	Num 8 6.
GQCONTACT	The name of the Group Quarters primary contact person.	Universe File	Char $140.
GQCURRMAXPOP	Maximum number of people at the Group Quarters.	Universe File, Master Address File	Num 8 6.
GQCURRSIZE	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Universe File, Master Address File	Num 8 6.
GQHUFLAG	Group Quarters/HU Flag. Used to identify the type of living quarter.	Universe File	Char $4.
GQNAME	The name of the Group Quarters, not the facility name.	Universe File	Char $400.
GQPER	Number of people at the Group Quarters MAFID	DRF1	Num 8
GQTYPCUR	Current Group Quarters Type code.	Universe File	Char $12.
MAFID	Master Address File ID - Permanent MAFUNIT ID	Universe File	Num 8 10.
MAF_CNT_0917	Count of sources (i.e more than one school) for Off Campus Persons on 9/17		Num 8
MAF_CNT_0924	Count of sources (i.e more than one school) for Off Campus Persons on 9/17		Num 8
MT1_SCC13	Indicator for more than one response through eResponse	DRF1	Num 8
REPUNIT_SPONSOR_CASE_ID	The unique identifier of the reporting unit. The MAFID is embedded in this variable.	DRF1	Char $12.
RESSTAT	Residential Status Flag	Universe File	Char $4.

SCC13	Number of people in Session_context_code 13, GQ eResponse	DRF1	Num	8
SCC13GP	Number of good people from Session_context_code 13, GQ eResponse.	DRF1	Num	8
SCC15	Number of people from Session_context_code 15, Individual Census Questionnaires	DRF1	Num	8
SCC15GP	Number of good people (GP=1) from session_context_code 15, Individual Census Questionnaires.	DRF1	Num	8
SCC17	Number of people from Session_context_code 17, GQ Paper Listings	DRF1	Num	8
SCC17GP	Number of good people (GP=1) from session_context_code 17, GQ Paper Listings.	DRF1	Num	8
SCC19	Number of people from Session_context_code 19, GQ Maritime Enumeration	DRF1	Num	8
SCC19GP	Number of good people (GP=1) from session_context_code 19, GQ Maritime Enumeration.	DRF1	Num	8
SCC22	Number of people from Session_context_code 22, Administrative Record (AR) Processing	DRF1	Num	8
SCC22GP	Number of good people (GP=1) from session_context_code 22, Administrative Record (AR) Processing.	DRF1	Num	8
SCC77	Number of people from Session_context_code 77, Post-Collection Processing Dummy Return created for GQ return not received by DRPS.	DRF1	Num	8
SCC77GP	Number of good people (GP=1) from session_context_code 77, Post-Collection Processing Dummy Return created for GQ return not received by DRPS.	DRF1	Num	8
SCC99	Number of people from Session_context_code 99, Catch-all for HU to GQ conversions indicating non-GQ SCCs.	DRF1	Num	8
SCC99GP	Number of good people (GP=1) from session_context_code 99, Catch-all for HU to GQ conversions indicating non-GQ SCCs.	DRF1	Num	8
UGP15_LATE	"unlinked" ICQs, where we have now done the linking; as well as any late ICQs received.		Num	8

Variables included in the Group Quarters File (GQ_MAFID_DSSD_OUT) provided by DSSD

Ryan King, Maranda Pepe, Hannah Zimmerman, Mary Frances Zelenak, Andy Keller, Juli Zamora

December 23, 2020

Variable	Definition	Source	Variable Type and Length
ACOCE	Area Census Office	Universe File	Char $16.
BCUCOUNTYFP	FIPS County Code	Universe File	Char $12.
BCUSTATEFP	FIPS State Code	Universe File	Char $8.
CDLPER	Indicator for Persons found in CDL, on 10/27.	CDL	Num 8
CDLPER_LATE	Indicator for Persons Found in CDL run on 12/8/20.	CDL	Num 8
DDP	Indicator for data-defined person.	DRF1	Num 8
DDP_PSA	Approximate number of data-defined people after preliminary DRF2 processing to resolve cases with more than one response at a MAFID.	** Post-DRF2 Processing Test File	Num 8
DRPS_CDL	Person records sent to both DRPS and CDL.	DRF1, CDL	Char 1
ER_STATED_CNT	The number of people residing in the Group Quarters on Census day, as stated by the Group Quarters contact person (administrator). Recorded during eResponse.	Case-Level File	Num 8 6.
FACTLNAME	The name of the Group Quarters facility. A facility is an umbrella organization that has a group of GQs. For example, a college is the facility and its dorms are the GQs.	Universe File	Char $400.
FOCS_CALCULATED	The number of people residing in the Group Quarters on Census day determined by the number of person responses recorded in FOCS.	Case-Level File	Num 8 6.
FOCS_ER_CB_CODE	A code that identifies the FOCS disposition status of the enumeration record associated with the Group Quarters.	Case-Level File	Char $12.
FOCS_STATED_CNT	The number of people residing in the Group Quarters on Census day, as stated by the Group Quarters contact person (administrator). Recorded in FOCS.	Case-Level File	Num 8 6.
GP	Number of people with good person flag, GP = 1	DRF1	Num 8
GP_PSA	Approximate number of good people (GP=1) after preliminary DRF2 processing to resolve cases with more than one response at a MAFID.	** Post-DRF2 Processing Test File	Num 8
GP13	Number of CDL people from Session_context_code 13, eResponse.	CDL	Num 8
GP15	Number of CDL people from Session_context_code 15, Individual Census Questionnaires	CDL	Num 8
GP17	Number of CDL peoplefrom Session_context_code 17, Paper Listings.	CDL	Num 8

GP13_LATE	all versions of the previous CDL indicators	CDL	Num 8
GP15_LATE	all versions of the previous CDL indicators	CDL	Num 8
UGP15_LATE	Number of unlinked ICQs from CDL, there is overlap with GP15_LATE.	CDL	Num 8
GP17_LATE	all versions of the previous CDL indicators	CDL	Num 8
GQ_OC_CNT_0917	Counts from GQ Off-Campus work on 9/17	Off-Campus Records	Num 8
GQ_OC_CNT_0924	Counts from GQ Off-Campus work on 9/24	Off-Campus Records	Num 8
GQ_PRI_PH_AREA_ID	The unique identifier of the phone number area code of the Group Quarters primary point of contact.	GQ Advanced Contact (GQAC)	Char $3.
GQ_PRI_PH_EXT_TEXT	The phone number extension of the Group Quarters primary point of contact.	GQ Advanced Contact (GQAC)	Char $8
GQ_PRI_PH_LINE_TEXT	The phone number line of the Group Quarters primary point of contact.	GQ Advanced Contact (GQAC)	Char $7.
GQ_SIZE_EXP_PERS_CNT	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact (GQAC)	Num 8 6.
GQ_SIZE_MAX_PERS_CNT	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact (GQAC)	Num 8 6.
GQCONTACT	The name of the Group Quarters primary contact person.	Universe File	Char $140.
GQCURRMAXPOP	Maximum number of people at the Group Quarters.	Universe File, Master Address File	Num 8 6.
GQCURRSIZE	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Universe File, Master Address File	Num 8 6.
GQHUFLAG	Group Quarters/HU Flag. Used to identify the type of living quarter.	Universe File	Char $4.
GQNAME	The name of the Group Quarters, not the facility name.	Universe File	Char $400.
GQPER	Number of people at the Group Quarters MAFID	DRF1	Num 8
GQTYPCUR	Current Group Quarters Type code.	Universe File	Char $12.
MAFID	Master Address File ID - Permanent MAFUNIT ID	Universe File	Num 8 10.
MAF_CNT_0917	Count of sources (i.e more than one school) for Off Campus Persons on 9/17		Num 8
MAF_CNT_0924	Count of sources (i.e more than one school) for Off Campus Persons on 9/17		Num 8
MT1_SCC13	Indicator for more than one response through eResponse	DRF1	Num 8
REPUNIT_SPONSOR_CASE_ID	The unique identifier of the reporting unit. The MAFID is embedded in this variable.	DRF1	Char $12.
RESSTAT	Residential Status Flag	Universe File	Char $4.

SCC13	Number of people in Session_context_code 13, GQ eResponse	DRF1	Num 8
SCC13GP	Number of good people from Session_context_code 13, GQ eResponse.	DRF1	Num 8
SCC15	Number of people from Session_context_code 15, Individual Census Questionnaires	DRF1	Num 8
SCC15GP	Number of good people (GP=1) from session_context_code 15, Individual Census Questionnaires.	DRF1	Num 8
SCC17	Number of people from Session_context_code 17, GQ Paper Listings	DRF1	Num 8
SCC17GP	Number of good people (GP=1) from session_context_code 17, GQ Paper Listings.	DRF1	Num 8
SCC19	Number of people from Session_context_code 19, GQ Maritime Enumeration	DRF1	Num 8
SCC19GP	Number of good people (GP=1) from session_context_code 19, GQ Maritime Enumeration.	DRF1	Num 8
SCC22	Number of people from Session_context_code 22, Administrative Record (AR) Processing	DRF1	Num 8
SCC22GP	Number of good people (GP=1) from session_context_code 22, Administrative Record (AR) Processing.	DRF1	Num 8
SCC77	Number of people from Session_context_code 77, Post-Collection Processing Dummy Return created for GQ return not received by DRPS.	DRF1	Num 8
SCC77GP	Number of good people (GP=1) from session_context_code 77, Post-Collection Processing Dummy Return created for GQ return not received by DRPS.	DRF1	Num 8
SCC99	Number of people from Session_context_code 99, Catch-all for HU to GQ conversions indicating non-GQ SCCs.	DRF1	Num 8
SCC99GP	Number of good people (GP=1) from session_context_code 99, Catch-all for HU to GQ conversions indicating non-GQ SCCs.	DRF1	Num 8
UGP15_LATE	"unlinked" ICQs, where we have now done the linking; as well as any late ICQs received.		Num 8

FLAGA	Flag for editing ratio (GP/ GQ_SIZE_EXP_PERS_CNT)
	M = ratio is missing
	R = review
	S = suppress from imputation base
	I = impute
		DSSD HB Edit	Char $1.

FLAGB	Flag for editing ratio (GP/ GQ_SIZE_MAX_PERS_CNT)
	M = ratio is missing
	R = review
	S = suppress from imputation base
	I = impute
		DSSD HB Edit	Char $1.

FLAGC	Flag for editing ratio (GP/GQCURRMAXPOP)
	M = ratio is missing
	R = review
	S = suppress from imputation base
	I = impute
		DSSD HB Edit	Char $1.

FLAGD	Flag for editing ratio (GP/GQCURRSIZE)
	M = ratio is missing
	R = review
	S = suppress from imputation base
	I = impute
		DSSD HB Edit	Char $1.

UNRES	Flag to indicate case is unresolved either due to zero pop or implausible pop	DSSD HB Edit	Num 8
EXPRATIO	Ratio (GP/ GQ_SIZE_EXP_PERS_CNT) for the nation	DSSD GQCI	Num 8
EXPRATIO_GQ	Ratio (GP/ GQ_SIZE_EXP_PERS_CNT) for the GQTYPCUR	DSSD GQCI	Num 8
EXPRATIO_GQ_ST	Ratio (GP/ GQ_SIZE_EXP_PERS_CNT) for the GQTYPCUR and BCUSTATEFP	DSSD GQCI	Num 8
IMP_RAT_EXP	Imputed count based on EXPRATIO	DSSD GQCI	Num 8
IMP_RAT_EXP_GQ	Imputed count based on EXPRATIO_GQ	DSSD GQCI	Num 8
IMP_RAT_EXP_GQ_ST	Imputed count based on EXPRATIO_GQ_ST	DSSD GQCI	Num 8
MAXRATIO	Ratio (GP/ GQ_SIZE_MAX_PERS_CNT) for the nation	DSSD GQCI	Num 8
MAXRATIO_GQ	Ratio (GP/ GQ_SIZE_MAX_PERS_CNT) for the GQTYPCUR	DSSD GQCI	Num 8
MAXRATIO_GQ_ST	Ratio (GP/ GQ_SIZE_MAX_PERS_CNT) for the GQTYPCUR and BCUSTATEFP	DSSD GQCI	Num 8
IMP_RAT_MAX	Imputed count based on MAXRATIO	DSSD GQCI	Num 8
IMP_RAT_MAX_GQ	Imputed count based on MAXRATIO_GQ	DSSD GQCI	Num 8
IMP_RAT_MAX_GQ_ST	Imputed count based on MAXRATIO_GQ_ST	DSSD GQCI	Num 8
CURRRATIO	Ratio (GP/GQCURRSIZE) for the nation	DSSD GQCI	Num 8
CURRRATIO_GQ	Ratio (GP/GQCURRSIZE) for the GQTYPCUR	DSSD GQCI	Num 8
CURRRATIO_GQ_ST	Ratio (GP/GQCURRSIZE) for the GQTYPCUR and BCUSTATEFP	DSSD GQCI	Num 8
IMP_RAT_CURR	Imputed count based on CURRRATIO	DSSD GQCI	Num 8
IMP_RAT_CURR_GQ	Imputed count based on CURRRATIO_GQ	DSSD GQCI	Num 8
IMP_RAT_CURR_GQ_ST	Imputed count based on CURRRATIO_GQ_ST	DSSD GQCI	Num 8
MAXCURRRATIO	Ratio (GP/GQCURRMAXPOP) for the nation	DSSD GQCI	Num 8
MAXCURRRATIO_GQ	Ratio (GP/GQCURRMAXPOP) for the GQTYPCUR	DSSD GQCI	Num 8
MAXCURRRATIO_GQ_ST	Ratio (GP/GQCURRMAXPOP) for the GQTYPCUR and BCUSTATEFP	DSSD GQCI	Num 8

IMP_RAT_CURRMAX	Imputed count based on MAXCURRRATIO	DSSD GQCI	Num 8
IMP_RAT_CURRMAX_GQ	Imputed count based on MAXCURRRATIO_GQ	DSSD GQCI	Num 8
IMP_RAT_CURRMAX_GQ_ST	Imputed count based on MAXCURRRATIO_GQ_ST	DSSD GQCI	Num 8
MEDGP	65th percentile of GP for the nation	DSSD GQCI	Num 8
MEDGP_GQ	65th percentile of GP by GQTYPCUR	DSSD GQCI	Num 8
MEDGP_GQ_ST	65th percentile of GP by GQTYPCUR and BCUSTATEFP. If GQTYPCUR = 104, 801, 802, and 901, 70th percentile If GQTYPCUR = 501, 68th percentile If GQTYPCUR = 301, 55th percentile	DSSD GQCI	Num 8
IMP_GP	Final Imputed Count	DSSD GQCI	Num 8
IMP_FLAG	Path Flag for Final Imputed Count.	DSSD GQCI	Num 8

For Consideration: An Alternate Approach to Model Selection and Validation for Group Quarters Count Imputation (GQCI)

12/21/20

Section 1. Defining the Models

Consider 7 models for GQCI:

M1a	Ratio method, using variable Va, GQAC expected count (requires Va)
M1b	Ratio method, using variable Vb, GQAC max count (requires Vb)
M1c	Ratio method, using variable Vc, current survey count (requires Vc)
M1d	Ratio method, using variable Vd, current survey max (requires Vd)
M2	Poisson model (requires Va, Vb, Vc, and Vd)
M3	Percentile method (requires none of these variables)
M4	CES method (~~different set of requirements~~requires none of these variables, but uses Va or Vd if available; Requires IPEDS room capacity (Ve) and Greek indicator (Vf))

Formatted: Space After: 0 pt

Formatted: Indent: Left: 2"

Basic approach to validating and comparing models. Splitting the set of good cases in the truth deck into 10 parts, run all seven models, and evaluate them within each Group Quarters (GQ) type, GQTi, i = 1, 2, ..., k. For each GQ type, determine the optimal ordering of the 7 models.

If some of the models can be eliminated, do so. If they are beaten by some available model uniformly across all GQ types (or nearly so), there is no need to pursue them.

Example. Suppose that, for GQTi, the order is {M1a, M1b, M2, M4, M3, M1d, M1c}. In each state, the procedure is as follows:

- If we have the GQAC expected count, apply M1a.
- If not, if we have the GQAC max count, apply M1b.
- (M2 is moot here, as we need all four variables. If we had had all four, we would have applied M1a.)
- If not, apply M4 if possible. (~~I don't know all of the CES requirements for model M4.~~requires Greek house indicator—always available; uses Va, Vd and IPEDS room capacity when available—IPEDS room cap is available for 98% of 501 GQs)
- If not, apply M3. NO NEED TO CONTINUE BEYOND M3 (because it's always available).

Similarly, for any sequence.

Note 1. The Poisson model, M2, is used only if it beats all other models within a GQ type.

Note 2. The optimal ordering of the models specific to a given GQ type can be determined through the validation of the seven models and then applied to each unresolved GQ, as described above. However, my recommendation is to simplify the process for model validation, as described in Section 2. This would lead to a simplification of the optimal ordering algorithm, as seen below.

Statistical advantage. With this approach, for each GQ within each GQ type, if the best model and variable(s) are available, we will use that model. If not, we go to the next best.

When writing a specification, we specify (1) the variable requirements for each model; if the variable(s) are not available, go to the next model; and (2) the optimal sequence of models (hierarchy) for each GQ type.

Section 2. Evaluating the Models

In this section, we consider the difficulty comparing models that have different requirements, including (1) a need for a different set of auxiliary variables, or (2) the availability of an outside set of data, such as the IPEDS, which provides GQ-facility level statistics for universities. Requirement (1) applies directly and easily to the first six models in Section 1. Requirement (2) pertains to M4, the CES method. Based on these considerations--and the reality of our time limit to incorporate a procedure for count imputation into production--we propose a simplified approach to evaluate the models and determine the details of the procedure.

In Step 1, we consider only the first six models in Section 1. Unlike the CES approach, these models are related in that each can be applied to any individual GQ without regard to its connection to a GQ facility. For the six models, the set of requirements involves the availability of some subset of the variables *V* = {Va, Vb, Vc, Vd}. To compare the performance of the six models within each GQ type, I propose that we restrict the truth deck of resolved, non-outlying GQs to the set for which we have all four variables. We can then perform a 10-fold model validation, derive the relevant prediction-error metrics, and compare fairly the imputation results on this set of cases, where each model is applicable.

Within a specific GQ type, this comparison would allow us to select the optimal ordering of the six models to apply when we have all four variables to impute an unresolved GQ. *We could then maintain this ordering regardless of the subset of V we have for any GQ*, that is, when we have only Va and Vc, none of them, etc. We run through the hierarchy until we have the variables we need for the model.

Alternatively--but not my recommendation--we could conduct a more thorough validation. We could divide the good cases in the truth deck into a number of subsets of the GQ universe. The first subset, $S_{4,1}$, would be that described above, in which every GQ has all four variables in *V*. We could then define a subset of GQs, $S_{3,1}$, for which each GQ has the variables Va, Vb, and Vc. A model validation on this subset would allow us to measure the performance of models M1a, M1b, M1c, and M3 when these three variables are available. Based on the results, in this circumstance, we could specify the preferred model from the four eligible models. In a similar manner, we could define additional subsets of GQs, $S_{3,2}$, $S_{3,3}$, and $S_{3,4}$, that have three of the four variables in *V*, and run model validations on each, producing the preferred model under those conditions.

We would then continue by defining six subsets of GQs in the truth deck, $S_{2,1}$, $S_{2,2}$, ..., $S_{2,6}$, for which the GQs have a specified two of the four variables available. For example, for the subset $S_{2,1}$, in which only Va and Vb are available, the model validation would allow a comparison of models M1a, M1b, and M3. We could continue in this way for the other five subsets of GQs that have only two variables in *V*. Finally, the four subsets $S_{1,1}$, $S_{1,2}$, $S_{1,3}$, and $S_{1,4}$ would allow us to compare each of the first four models individually to M3 for the application when only one variable is available.

Rather than conduct 15 model validations as described here, I believe and hope we would obtain similar results much more quickly and easily by conducting only the first validation described above, and then spending the time to thoroughly analyze its results. This would allow us to set an optimal ordering of the first six models within each GQ type.

In Step 2, we integrate the CES model, M4, into the sequence for any M4-applicable GQ type, that is, for which M4 can be applied, such as colleges and nursing homes. For such a GQ type, define the truth deck as those resolved GQs for which M4 can be applied. Within this GQ type GQTi, the optimal ordering of the first 6 models can be based on the results of the model validation in Step 1. Define Model $M5_{GQTi}$ as the procedure that applies this ordering of the models. The subscript is inserted to indicate that M5 will likely differ across GQ types.

Next, we conduct a model validation that compares M4 to $M5_{GQTi}$. By using the set of GQs for which M4 can be applied, we can compare its performance fairly to that of $M5_{GQTi}$. If M4 performs better than $M5_{GQTi}$ for any GQ type, it should be placed first in the optimal ordering for that GQ type. That is, it should be applied when the data requirements for M4 are satisfied. If $M5_{GQTi}$ performs better than M4, M4 can be eliminated from consideration for that GQ type.

To be copied/edited into "Group Quarters Imputation Methodology"

Possible Methods

The GQ count imputation will be hierarchical, following ~~three~~ four steps:

1. Substitution
2. Modeling
3. Sort or Mean
4. Hierarchical substitution with adjusted residual

Adjusted Residual from Facility-level Total

This method can be used as a complement to the Substitution method. After first imputing the reported count with the GQ advance contact expected count (if available), we will implement the following facility-level residual method. This method can only be used for GQs with GQTYPCUR=501.

For universities and colleges, we have the 2019 facility-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the facility level to the GQ data. We will adjust the room capacity for GQ population in off-campus Greek housing (which is not included in the IPEDS room capacity). To avoid overcounting, we will also scale the room capacity by the average ratio (within facility size classes) of the facility-level total 2020 Census Day population over the facility-level room capacity. For calculating these ratios, we will only use facilities with for which less than 5 or 10% of the GQs at the facility are unresolved cases.

Commented [TKW(F1]: Joe Staudt can provide a description of the matching algorithm, and the quality of the matches (which is very high for a high percentage of the cases).

How to portion out the residual to multiple GQs

For facilities with more than one unresolved GQ, we will need to impute the fraction of the facility-level residual population that goes with each GQ. We propose a hierarchy of two approaches:

1. We will sum the reported GQ population counts from the 2010 Decennial to facility level. (This data has already been merged on mafid to the 2020 GQ counts file. Then for each GQ we will calculate its share of the facility's population. For GQs (mafid) that existed in 2010 and still exist in 2020 we will these 2010 GQ shares of facility-level population to calculate the share of the facility's residual population (calculated as described above) at each unresolved GQ. For any unresolved GQs that cannot be imputed this way, we will follow approach 2.
2. For unresolved GQs that did not exist in 2010 (and for which we have no 2020 GQ-level estimate), we will divide the residual facility-level population evenly among the remaining GQs.

To be copied/edited into "Group Quarters Imputation Methodology"

Possible Methods

The GQ count imputation will be hierarchical, following ~~three~~ four steps:

1. Substitution
2. Modeling
3. Sort or Mean
4. Hierarchical substitution with adjusted residual

Adjusted Residual from Facility-level Total

This method can be used as a complement to the Substitution method. After first imputing the reported count with the GQ advance contact expected count (if available), we will implement the following facility-level residual method. This method can only be used for GQs with GQTYPCUR=501.

For universities and colleges, we have the 2019 facility-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the facility level to the GQ data. We will adjust the room capacity for GQ population in off-campus Greek housing (which is not included in the IPEDS room capacity). To avoid overcounting, we will also scale the room capacity by the average ratio (within facility size classes) of the facility-level total 2020 Census Day population over the facility-level room capacity. For calculating these ratios, we will only use facilities with for which less than 5 or 10% of the GQs at the facility are unresolved cases.

Commented [TKW(F1]: Joe Staudt can provide a description of the matching algorithm, and the quality of the matches (which is very high for a high percentage of the cases).

Algorithm for Adjusting the residual

To adjust the room capacity residual, we want a set of GQs for which we can estimate the average utilization rate of the GQs. For example, if a dorm has enough rooms for 100 persons, on average how many people do we think would actually be living in the dorm on Census Day. To estimate this, we first select facilities (universities) for which we have a positive GQAC Max Number of People (GQCURRMAXPOP) for every GQ at the facility. Since the IPEDS data does not include off-campus housing, we further subset on facilities that have no Greek GQs. Finally, to maximize the chances that we are comparing apples to apples, we also subset to facilities for which the match quality is very high (match score > 90%). Within this subset, we calculate the average ratio of the facility-level sum of GQAC Max Number of People over the room capacity from IPEDS:

$$Average\ Ratio_S = \sum_{i \in S} \frac{\sum_{facility\_i} GQAC\ Max\ Number\ of\ People}{IPEDS\ Room\ Capacity\ at\ facility\ i}$$

where S is the set of facilities with no Greek GQs only positive values for GQAC Max Number of People.

How to portion out the residual to multiple GQs

For facilities with more than one unresolved GQ, we will need to impute the fraction of the facility-level residual population that goes with each GQ. We propose a hierarchy of two approaches:

1. We will sum the reported GQ population counts from the 2010 Decennial to facility level. (This data has already been merged on mafid to the 2020 GQ counts file. Then for each GQ we will calculate its share of the facility's population. For GQs (mafid) that existed in 2010 and still exist in 2020 we will these 2010 GQ shares of facility-level population to calculate the share of the facility's residual population (calculated as described above) at each unresolved GQ. For any unresolved GQs that cannot be imputed this way, we will follow approach 2.
2. For unresolved GQs that did not exist in 2010 (and for which we have no 2020 GQ-level estimate), we will divide the residual facility-level population evenly among the remaining GQs.

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. A telephone operation is in progress to collect data for these GQs. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that were vacant during GQ Advanced Contact (GQAC) but were open on Census Day require imputation.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Vision but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported.

Commented [PJC(F1]: For condition situation (2), do we have to mention other statuses?

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Commented [PJC(F2]: We should explain, perhaps insert a note below the table or in the text that differentiates "Vacant During Visit, Open on Census Day" from "Vacant GQ"? I presume the two are mutually exclusive. So did the former report that they were open on CD, while the latter didn't report anything about CD? I don't understand.

Table 2 shows the status of the occupied GQs. There are 43,000 unresolved occupied GQs and 184,000 resolved occupied GQs.

Table 2: Occupied GQ Universe

Commented [JEZ(F3]: Needs to be updated with latest from Ryan.

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Total	184,000	43,000	227,000

The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs. Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 6 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

GQ Type	Census Day Pop	Unresolved	Total
Correctional Facilities	13,000	2,800	16,000
Juvenile Facilities	6,200	1,800	8,000

Nursing Facilities	25,500	3,200	28,500
Hospitals	2,000	800	2,800
College Housing	30,500	5,600	36,000
Military	3,100	1,900	5,000
Shelters	24,500	8,200	33,000
Group Homes	62,500	9,100	72,000
Other	16,000	9,700	26,000
Total	184,000	43,000	227,000

Imputation Methods

Variables

Table 4 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include the current, 2020 Decennial Response File 1(DRF 1), the 2010 CUF, Master Address File, Administrative Records, GQ Advanced Contacts, the American Community Survey, data collected via web-scraping, data from the IPEDS, and data from the Common Core. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Commented [PJC(F4]: I'd move this to second, after 2020 DRF1.

Table 4: Auxiliary and Historical Data

Commented [PJC(F5]: Should we split this table into two parts? This part would be GQ-Level Variables, the second would be GQ-Facility-Level Variables? This would point it out to the reader.

Variable	Description	Source
GQ Type	Based on first digit of GQTYPCUR	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR
Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQ Advanced Contact	Master Address File / DRF1
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
[Web Scraping Vars]		
Room Cap	Description of ROOMCAP: If institution provides on-campus housing, specify dormitory capacity for academic year DORMITORY CAPACITY - The maximum number of students that the institution can provide residential facilities for, whether on or off campus. (off-campus dormitory space that is reserved by the institution).	IPEDS

Commented [JEZ(F6]: Need CES to provide details

Commented [JEZ(F7]: Need CES to provide details

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

The GQ count imputation will be hierarchical, following three steps:

1. Substitution
2. Modeling
3. Sort or Mean

Substitution

First i~~I~~f a pop count is available from ~~another 2020 Census source, such as from~~ the NPC call operation ~~or the GQ advance contact~~, we will substitute ~~with~~ that pop count. If not for cases where we have an expected GQ pop count or a maximum GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will substitute a function of those variables. Although the expected GQ count from the GQAC ~~advanced contact i~~was not ~~a~~ reported during the GQ Enumeration (GQE), ~~count for the 2020 Census, we will first impute with the GQ advance contact expected count, if available~~we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during ~~-~~GQE should provide information on this presumption and on functions of the expected GQ pop count and the maximum GQ pop count that produce more accurate imputation.

Table 5 shows that 8,600 of the unresolved cases can be resolved by substituting with the GQAC expected count.

We will not substitute with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count.

Table 5: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Modeling

For the unresolved GQs without a GQ advance contact expected count, we will attempt to substitute a GQ pop count from the American Community Survey or Census 2010. If no GQ pop can be found for the unresolved GQ and sufficient auxiliary variables are available, we will impute with a prediction from a logistic or Poisson regression model. For the logistic regression the dependent variable will be the reported count / Max GQ size. For the Poisson regression, the dependent variable will be reported GQ pop count with an offset of the current GQ size (because it is the most often filled size variable). Independent variables will be selected from Table 4. It is important to note that GQ type will either be a covariate in the models or separate models will be fit by GQ type. Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

In the event that models cannot be built or implemented in time, we will use the ratio imputation method. For the ratio imputation method, we will group the GQs by type and calculate the ratio of the sum of the auxiliary variable to the sum of the reported GQ population. For the unresolved GQs, we will multiply the auxiliary variable by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = \ GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

Commented [JEZ(F8]: I implemented this method for Max Number of People, Current GQ Size, GQAC Max Number of People, and GQAC Expected Count. I can create some metrics tomorrow.

Commented [TLK(F9R8]: I think this should be very similar to running a Poisson regression with an intercept.

Sort or Mean

If sufficient auxiliary data is not available, we will either a) find a donor with the most matching characteristics as the unresolved unit or b) impute the pop size with average population within an imputation cell. Both methods involve partitioning the GQ universe into imputation cells based on GQ type. To find the most similar donor, the GQs will be sorted with GQ type by

- Max Number of People
- Current GQ Size
- AR Count
- 2010 GQ Count
- Detailed GQ type
- State
- County
- BCU
- MAFID

Then the GQ pop size of the previous resolved occupied GQ in the sort will be carried over into the unresolved GQ. This sort is selected so that GQs with similar sizes are sorted together.

Alternatively, we could form cells based on the Maximum Number of People (modulo 10) and detailed GQ type. Then, we will calculated the average GQ population size and impute the unresolved GQs with the average.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

Models

We will evaluated the imputation models using cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The reaming 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will predict the GQ pop size for the validation deck. For the validation deck, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each GQ. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Sort or Mean

To evaluate the sort and mean methods, we will simply conduct the procedures and then review them for reasonableness.

Appendices

Table 6: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Table 7: Resolved and Unresolved Counts by GQ Type

GQ Type	Resolved	Unresolved	Total
Correctional Facilities	13,000	2,800	16,000
Juvenile Facilities	6,200	1,800	8,000
Nursing Facilities	25,500	3,200	28,500
Hospitals	2,000	800	2,800
College Housing	30,500	5,600	36,000
Military	3,100	1,900	5,000
Shelters	24,500	8,200	33,000
Group Homes	62,500	9,100	72,000
Other	16,000	9,700	26,000
Total	184,000	43,000	227,000

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

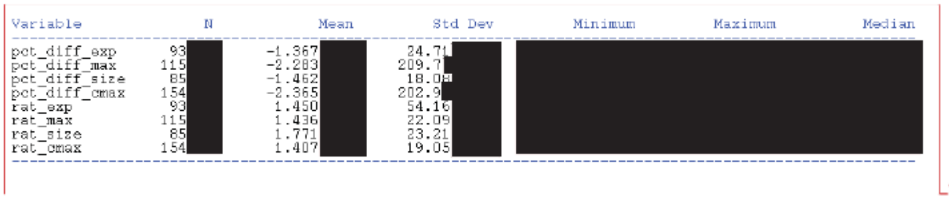
Current GQ Size	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

Max Number of People	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

GQ Type	Ratio of GQAC Expected Count to Good People Count	Ratio of GQAC Max Number of People to Good People Count	Ratio of Current GQ Size to Good People Count	Ratio of Max Number of People to Good People Count
Correctional Facilities	0.7239	0.4350	0.9223	0.4468
Juvenile Facilities	0.6977	0.3075	0.8621	0 3284
Nursing Facilities	0.8875	0.6815	0.9705	0.6810
Hospitals	0.7779	0.6424	1 017	0.6441
College Housing	0.8203	0.6147	1 071	0.6114

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Military	0.7533	0.2347	1.047	0 3326
Shelters	0.6875	0.5768	0.6762	0.6070
Group Homes	0.8866	0.5374	1.048	0 5307
Other	0.8283	0.4205	1.084	0.4002
All GQs	0.8211	0.5377	0.9851	0 5478



```
Variable              N              Mean         Std Dev          Minimum          Maximum            Median
--------------------------------------------------------------------------------------------------------------
pct_diff_exp         93[redacted]   -1.367[redacted]  24.71[redacted]  [redacted]
pct_diff_max        115             -2.283           209.7
pct_diff_size        85             -1.462           18.0
pct_diff_cmax       154             -2.365           202.9
rat_exp              93              1.450           54.16
rat_max             115              1.436           22.09
rat_size             85              1.771           23.21
rat_cmax            154              1.407           19.05
--------------------------------------------------------------------------------------------------------------
```

Commented [JEZ(F10]: I can format this tomorrow. Key: pct_diff = (GP – Imputed value)/GP for resolved cases. Rat = GP/AUX. Exp = GQAC Expected Count
Max = GQAC Max Number of People
Size = Current GQ Size
CMax = Max Number of People

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. A telephone operation is in progress to collect data for these GQs. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that were vacant during GQ Advanced Contact (GQAC) but were open on Census Day require imputation.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but do not have a reported count. This universe is made up of GQs with a status of Occupied, Vacant During Vision but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported.

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Table 2 shows the status of the occupied GQs. There are 43,000 unresolved occupied GQs and 184,000 resolved occupied GQs.

Table 2: Occupied GQ Universe

Commented [JEZ(F1]: Needs to be updated with latest from Ryan.

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Total	184,000	43,000	227,000

The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs. Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 6 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

GQ Type	Census Day Pop	Unresolved	Total
Correctional Facilities	13,000	2,800	16,000
Juvenile Facilities	6,200	1,800	8,000
Nursing Facilities	25,500	3,200	28,500

Hospitals	2,000	800	2,800
College Housing	30,500	5,600	36,000
Military	3,100	1,900	5,000
Shelters	24,500	8,200	33,000
Group Homes	62,500	9,100	72,000
Other	16,000	9,700	26,000
Total	184,000	43,000	227,000

Imputation Methods

Variables

Table 4 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include the current, 2020 Decennial Response File 1(DRF 1), the 2010 CUF, Master Address File, Administrative Records, GQ Advanced Contacts, the American Community Survey, data collected via web-scraping, data from the IPEDS, and data from the Common Core. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 4: Auxiliary and Historical Data

Variable	Description	Source
GQ Type	Based on first digit of GQTYPCUR	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR
Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQ Advanced Contact	Master Address File / DRF1
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
[Web Scraping Vars]		
Room Cap	Description of ROOMCAP: If institution provides on-campus housing, specify dormitory capacity for academic year DORMITORY CAPACITY - The maximum number of students that the institution can provide residential facilities for, whether on or off campus. (off-campus dormitory space that is reserved by the institution).	IPEDS

Commented [JEZ(F2]: Need CES to provide details

Commented [JEZ(F3]: Need CES to provide details

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

The GQ count imputation will be hierarchical, following three steps:

1. Substitution
2. Modeling
3. Sort or Mean

Substitution

If a pop count is available from another 2020 Census source, such as from the NPC call operation or the GQ advance contact, we will substitute with that pop count. Although the expected GQ count from the GQ advanced contact is not a reported count for the 2020 Census, we will first impute with the GQ advance contact expected count, if available. Table 5 shows that 8,600 of the unresolved cases can be resolved by substituting with the GQAC expected count.

We will not substitute with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count.

Table 5: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Modeling

For the unresolved GQs without a GQ advance contact expected count, we will attempt to substitute a GQ pop count from the American Community Survey or Census 2010. If no GQ pop can be found for the unresolved GQ and sufficient auxiliary variables are available, we will impute with a prediction from a logistic or Poisson regression model. For the logistic regression the dependent variable will be the reported count / Max GQ size. For the Poisson regression, the dependent variable will be reported GQ pop count with an offset of the current GQ size (because it is the most often filled size variable). Independent variables will be selected from Table 4. It is important to note that GQ type will either be a covariate in the models or separate models will be fit by GQ type. Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

In the event that models cannot be built or implemented in time, we will use the ratio imputation method. For the ratio imputation method, we will group the GQs by type and calculate the ratio of the sum of the auxiliary variable to the sum of the reported GQ population. For the unresolved GQs, we will multiply the auxiliary variable by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = \ GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

Commented [JEZ(F4]: I implemented this method for Max Number of People, Current GQ Size, GQAC Max Number of People, and GQAC Expected Count. I can create some metrics tomorrow.

Commented [TLK(F5R4]: I think this should be very similar to running a Poisson regression with an intercept.

Sort or Mean

If sufficient auxiliary data is not available, we will either a) find a donor with the most matching characteristics as the unresolved unit or b) impute the pop size with average population within an imputation cell. Both methods involve partitioning the GQ universe into imputation cells based on GQ type. To find the most similar donor, the GQs will be sorted with GQ type by

- Max Number of People
- Current GQ Size
- AR Count
- 2010 GQ Count
- Detailed GQ type
- State
- County
- BCU
- MAFID

Then the GQ pop size of the previous resolved occupied GQ in the sort will be carried over into the unresolved GQ. This sort is selected so that GQs with similar sizes are sorted together.

Alternatively, we could form cells based on the Maximum Number of People (modulo 10) and detailed GQ type. Then, we will calculated the average GQ population size and impute the unresolved GQs with the average.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

Models

We will evaluated the imputation models using cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The reaming 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will predict the GQ pop size for the validation deck. For the validation deck, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each GQ. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Sort or Mean

To evaluate the sort and mean methods, we will simply conduct the procedures and then review them for reasonableness.

Appendices

Table 6: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Table 7: Resolved and Unresolved Counts by GQ Type

GQ Type	Resolved	Unresolved	Total
Correctional Facilities	13,000	2,800	16,000
Juvenile Facilities	6,200	1,800	8,000
Nursing Facilities	25,500	3,200	28,500
Hospitals	2,000	800	2,800
College Housing	30,500	5,600	36,000
Military	3,100	1,900	5,000
Shelters	24,500	8,200	33,000
Group Homes	62,500	9,100	72,000
Other	16,000	9,700	26,000
Total	184,000	43,000	227,000

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

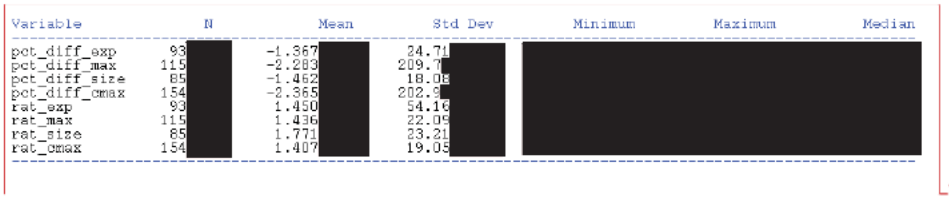
GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Current GQ Size	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

Max Number of People	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

GQ Type	Ratio of GQAC Expected Count to Good People Count	Ratio of GQAC Max Number of People to Good People Count	Ratio of Current GQ Size to Good People Count	Ratio of Max Number of People to Good People Count
Correctional Facilities	0.7239	0.4350	0.9223	0.4468
Juvenile Facilities	0.6977	0.3075	0.8621	0 3284
Nursing Facilities	0.8875	0.6815	0.9705	0.6810
Hospitals	0.7779	0.6424	1 017	0.6441
College Housing	0.8203	0.6147	1 071	0.6114

Military	0.7533	0.2347	1.047	0 3326
Shelters	0.6875	0.5768	0.6762	0.6070
Group Homes	0.8866	0.5374	1.048	0 5307
Other	0.8283	0.4205	1.084	0.4002
All GQs	0.8211	0.5377	0.9851	0 5478



```
Variable              N              Mean        Std Dev         Minimum        Maximum        Median
-----------------------------------------------------------------------------------------------------
pct_diff_exp         93 [redacted]  -1.367 [redacted]  24.71 [redacted]  [redacted]
pct_diff_max        115             -2.283             209.7
pct_diff_size        85             -1.462             18.08
pct_diff_cmax       154             -2.365             202.9
rat_exp              93              1.450             54.16
rat_max             115              1.436             22.09
rat_size             85              1.771             23.21
rat_cmax            154              1.407             19.05
-----------------------------------------------------------------------------------------------------
```

Commented [JEZ(F6]: I can format this tomorrow. Key: pct_diff = (GP – Imputed value)/GP for resolved cases. Rat = GP/AUX. Exp = GQAC Expected Count
Max = GQAC Max Number of People
Size = Current GQ Size
CMax = Max Number of People

Group Quarters Imputation Methodology

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. A telephone operation is in progress to collect data for these GQs. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that were vacant during GQ Advanced Contact (GQAC) but were open on Census Day require imputation.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but do not have a reported count. This universe is made up of GQs with a status of Occupied, Vacant During Vision but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported.

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Table 2 shows the status of the occupied GQs. There are 43,000 unresolved occupied GQs and 184,000 resolved occupied GQs.

Table 2: Occupied GQ Universe

Commented [JEZ(F1]: Needs to be updated with latest from Ryan.

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Total	184,000	43,000	227,000

The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs. Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 7 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

GQ Type	Census Day Pop	Unresolved	Total
Correctional Facilities	13,000	2,800	16,000

Juvenile Facilities	6,200	1,800	8,000
Nursing Facilities	25,500	3,200	28,500
Hospitals	2,000	800	2,800
College Housing	30,500	5,600	36,000
Military	3,100	1,900	5,000
Shelters	24,500	8,200	33,000
Group Homes	62,500	9,100	72,000
Other	16,000	9,700	26,000
Total	184,000	43,000	227,000

Imputation Methods

Variables

Table 4 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include the current, 2020 Decennial Response File 1(DRF 1), the 2010 CUF, Master Address File, Administrative Records, GQ Advanced Contacts, the American Community Survey, data collected via web-scraping, data from the IPEDS, and data from the Common Core. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 4: Auxiliary and Historical Data

Variable	Description	Source
GQ Type	Based on first digit of GQTYPCUR	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (0,1,2,3,4,5,6+)	AR
Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQ Advanced Contact	Master Address File / DRF1
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
[Web Scraping Vars]		
Room Cap		IPEDS

Commented [JEZ(F2]: Need CES to provide details

Commented [JEZ(F3]: Need CES to provide details

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

The GQ count imputation will be hierarchical. If a pop count is available from another 2020 Census source, such as from the NPC call operation or the GQ advance contact, we will substitute with that pop count. Although the expected GQ count from the GQ advanced contact is not a reported count for the 2020 Census, we will first impute with the GQ advance contact expected count, if available. Table 5 shows that 8,600 of the unresolved cases can be resolved by substituting with the GQAC expected count.

Table 5: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For the unresolved GQs without a GQ advance contact expected count, we will attempt to substitute a GQ pop count from the American Community Survey or Census 2010. If no GQ pop can be found for the unresolved GQ and sufficient auxiliary variables are available, we will impute with one of

If the GQ expected population count is not available, we will

We will develop a hierarchy of imputation methods, with conditions dependent upon the available data for a given unresolved GQ. We will have four(more?) possible types of imputation:

- Substitution
- Ratio Imputation
- Nearest Neighbor
- Poisson Regression

Commented [JEZ(F4]: Do we want to use this? What are we using for distance?

In the first method, we will directly substitute a given population count from historical or auxiliary data. For instance, if an expected count is available from the GQ Advanced Contact operation(?) we will use this value for the imputed population count. Possible contenders for substitution include the Current GQ Size, GQAC Expected Count, 2010 GQ Count and 2010 Occupied HU count.

For the ratio imputation method, we will group the GQs by type and calculate the ratio of the sum of the auxiliary variable to the sum of the reported GQ population. For the unresolved GQs, we will multiply the auxiliary variable by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

Commented [JEZ(F5]: I implemented this method for Max Number of People, Current GQ Size, GQAC Max Number of People, and GQAC Expected Count. I can create some metrics tomorrow.

Commented [TLK(F6R5]: I think this should be very similar to running a Poisson regression with an intercept.

We will fit nine Poisson regression models, one for each GQ type. The dependent variable is the good person count. Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators. We will test models with offsets based on the GQAC expected count and GQAC Maximum Number of People.

Question: Are there any other methods we should explore?

Hierarchy of Methods

We will determine a hierarchy of methods based on how the different methods perform when compared to the reported values.

Percent difference? Absolute difference?

Table 6: Example Imputation Method Hierarchy

Imputation Method	Condition(s)
Substitute 2010 GQ Count or 2010 Occupied HU Count	2010 GQ Count or 2010 Occupied HU Count available for MAFID.
[More methods?]	[Insert more business rules here? [e.g. GQ Type is Dorm]
Poisson Regression – Model 1	GQ Type is college
Poisson Regression – Model 2	GQ Type is not college
Ratio Methods	

Metrics

Poisson Models – Cross Validation

Appendices

Table 7: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles

CODE	VALUE
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Table 8: Resolved and Unresolved Counts by GQ Type

GQ Type	Resolved	Unresolved	Total
Correctional Facilities	13,000	2,800	16,000
Juvenile Facilities	6,200	1,800	8,000
Nursing Facilities	25,500	3,200	28,500
Hospitals	2,000	800	2,800
College Housing	30,500	5,600	36,000
Military	3,100	1,900	5,000
Shelters	24,500	8,200	33,000
Group Homes	62,500	9,100	72,000
Other	16,000	9,700	26,000
Total	184,000	43,000	227,000

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

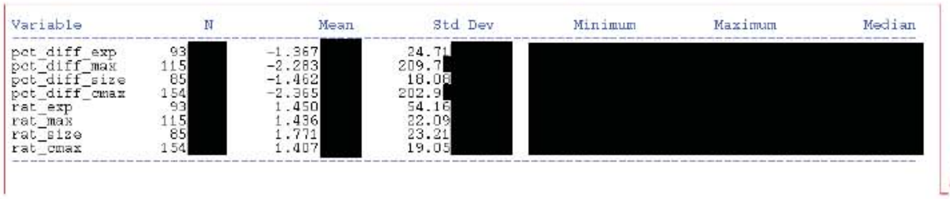
DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Current GQ Size	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

Max Number of People	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

GQ Type	Ratio of GQAC Expected Count to Good People Count	Ratio of GQAC Max Number of People to Good People Count	Ratio of Current GQ Size to Good People Count	Ratio of Max Number of People to Good People Count
Correctional Facilities	0.7239	0.4350	0.9223	0.4468
Juvenile Facilities	0.6977	0.3075	0.8621	0 3284
Nursing Facilities	0.8875	0.6815	0.9705	0.6810
Hospitals	0.7779	0.6424	1.017	0.6441
College Housing	0.8203	0.6147	1.071	0.6114
Military	0.7533	0.2347	1.047	0 3326
Shelters	0.6875	0.5768	0.6762	0.6070
Group Homes	0.8866	0.5374	1.048	0 5307
Other	0.8283	0.4205	1.084	0.4002
All GQs	0.8211	0.5377	0.9851	0 5478



```
Variable              N           Mean        Std Dev        Minimum        Maximum         Median
----------------------------------------------------------------------------------------------------
pct_diff_exp         93[redacted]  -1.367[redacted]  24.71[redacted]  [redacted]
pct_diff_max        115         -2.283         209.7
pct_diff_size        85         -1.462         18.0
pct_diff_cmax       154         -2.365         202.9
rat_exp              93          1.450         54.16
rat_max             115          1.436         22.09
rat_size             85          1.771         23.21
rat_cmax            154          1.407         19.05
----------------------------------------------------------------------------------------------------
```

Commented [JEZ(F7]: I can format this tomorrow. Key:
pct_diff = (GP – Imputed value)/GP for resolved cases. Rat = GP/AUX. Exp = GQAC Expected Count
Max = GQAC Max Number of People
Size = Current GQ Size
CMax = Max Number of People

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

Background

Commented [JEZ(F1]: Tables based on 12/13/20 data.

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A telephone operation is in progress to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that were vacant during GQ Advanced Contact (GQAC) but were open on Census Day require imputation.

In addition, we will impute a pop size for GQs that have a reported Census Day population count that is much smaller than expected. Our initial proposal is to impute when the Census Day population count is 25% of the GQAC expected count, but research into determining this threshould (and refining it) is ongoing.

Imputation Universe

Commented [JEZ(F2]: Meeting comments: Need to decide if we are imputing for low counts or only missing/zero counts. If we are going to impute for discrepancy cases, what are the conditions or thresholds?

Need to address which cases we will use treat as resolved for imputation – i.e. "donors".

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Visit but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with much lower than expected population count are included in the Census Day Pop column.

Commented [PJC(F3]: For condition situation (2), do we have to mention other statuses?

Commented [JEZ(F4R3]: I think this question needs SME input. Are we accepting Vacant, Delete and Nonresidential GQ counts as-is?

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Commented [PJC(F5]: We should explain, perhaps insert a note below the table or in the text that differentiates "Vacant During Visit, Open on Census Day" from "Vacant GQ"? I presume the two are mutually exclusive. So did the former report that they were open on CD, while the latter didn't report anything about CD? I don't understand.

Commented [JEZ(F6R5]: Need to ask Debbie and Ryan.

Table 2 shows the status of the occupied GQs. There are 43,000 unresolved occupied GQs without a census day population.

Table 2: Occupied GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000

Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Total	184,000	43,000	227,000

Additionally some of the 184,000 resolved occupied GQs will be treated as unresolved because their census day population is much lower than expected.

Commented [TLK(F7]: We need to calculate how many GQs have a Census Day population that is 25% of the GQAC expected count.

GQ Status	Reasonable Census Day Pop	Unresolved		Total
		Low Census Day Pop	No Census Day Pop	
Occupied GQ			17,000	197,000
Vacant During Visit, Open on Census Day			19,500	21,500
Refusal GQ			6,700	7,800
Total			43,000	227,000

The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs as well as any GQs with a much lower than expected population count. Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 7 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

GQ Type	Census Day Pop	Unresolved	Total
Correctional Facilities*	13,000	2,800	16,000
Juvenile Facilities	6,200	1,800	8,000
Nursing Facilities*	25,500	3,200	28,500
Hospitals	2,000	800	2,800
College Housing*	30,500	5,500	36,000
Military*	3,100	1,900	5,000
Shelters	24,500	8,200	33,000
Group Homes	62,500	9,100	72,000
Other	16,000	9,700	26,000
Total	184,000	43,000	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Variables

Table 4 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, and Administrative Records. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 4: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source

GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQ Advanced Contact	Master Address File / DRF1
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Vacant During Visit, Open on Census Day; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR

Additional sources available for college housing GQs include data collected via web-scraping, data from the Integrated Postsecondary Education Data System (IPEDS) and data from the Common Core. These variables are available at the facility level but not for individual MAFIDs. For universities and colleges, we have the 2019 facility-level total room capacity (number of persons that could live in the GQ) from the IPEDS. To obtain these data….[information about CES matching]. The room capacity variable is of high-quality and is available for most (95%) of college housing.

Commented [JEZ(F8]: Is this only for 501s?

Table 5: Facility-Level Data available for College Housing GQs

Variable	Description	Source
[Web Scraping Vars]		
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Commented [JEZ(F9]: Need CES to provide details

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

First, if a pop count is available from the NPC call operation, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will be hierarchical, following three steps:

1. Conversion from GQAC Expected Count
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling

4. Mean Imputation

Conversion from GQAC

For cases where we have an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will substitute a function of those variables. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 6 shows that 8,600 of the unresolved cases can be resolved by substituting with the GQAC expected count.

For each GQ type, we will use the ratio of the reported GQ Census Day count to the GQAC expected count to convert the GQAC expected count of the unresolved GQ to a Census Day imputed count. For each GQ type, we will calculate the ratio of the sum of the GQAC Expected Count to the sum of the reported GQ population for the resolved cases. For the unresolved GQs, we will multiply the GQAC expected count by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will not substitute with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count.

Table 6: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Adjusted Residual from Facility-level Total for College Housing

If the GQ advance contact expected count is not populated, we will implement the following facility-level residual method. This method can only be used for GQs with GQTYPCUR=501.

For universities and colleges, we have the 2019 facility-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the facility level to the GQ data. We will adjust the room capacity for GQ population in off-campus Greek housing (which is not included in the IPEDS room capacity). To avoid overcounting, we will also scale the room capacity by the average ratio (within facility size classes) of the facility-level total 2020 Census Day population over the facility-level room capacity. For calculating these ratios, we will only use facilities with for which less than 5 or 10% of the GQs at the facility are unresolved cases.

Commented [TKW(F10]: Joe Staudt can provide a description of the matching algorithm, and the quality of the matches (which is very high for a high percentage of the cases).

How to portion out the residual to multiple GQs

For facilities with more than one unresolved GQ, we will need to impute the fraction of the facility-level residual population that goes with each GQ. We propose a hierarchy of two approaches:

1. We will sum the reported GQ population counts from the 2010 Decennial to facility level. (This data has already been merged on mafid to the 2020 GQ counts file. Then for each GQ we will calculate its share of the facility's population. For GQs (mafid) that existed in 2010 and still exist in 2020 we will these 2010 GQ shares of facility-level population to calculate the share of the facility's residual population (calculated as described above) at each unresolved GQ. For any unresolved GQs that cannot be imputed this way, we will follow approach 2.
2. For unresolved GQs that did not exist in 2010 (and for which we have no 2020 GQ-level estimate), we will divide the residual facility-level population evenly among the remaining GQs.

Modeling

If the previous two steps do not yield an imputation (no GQAC expected count and no IPEDS count) for the unresolved GQ and sufficient auxiliary variables are available, we will impute with a prediction from a logistic or Poisson regression model. For the logistic regression the dependent variable will be the reported count / max number of people (because it is the most often filled size variable). For the Poisson regression, the dependent variable will be reported GQ pop count with an offset of the max number of people. Independent variables will be selected from Table 4. It is important to note that GQ type will either be a covariate in the models or separate models will be fit by GQ type. Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

Mean Imputation

If sufficient auxiliary data is not available, we will impute the pop size with average population within an imputation cell. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and GQ status. Then, we will calculate the average GQ population size and impute the unresolved GQs with the average.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

Models

We will evaluate the imputation models using cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The remaining 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will predict the GQ pop size for the validation deck. For the validation deck, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each GQ. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Conversion and Mean

To evaluate the Conversion from GQAC and mean methods, we will simply conduct the procedures and then review them for reasonableness.

Appendices

Table 7: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

GQ Type	Ratio of GQAC Expected Count to Good People Count	Ratio of GQAC Max Number of People to Good People Count	Ratio of Current GQ Size to Good People Count	Ratio of Max Number of People to Good People Count
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0 3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1 017	0.6385
College Housing	0.7818	0.5492	0.9444	0.5535
Military	0.7317	0.2290	0.9492	0.2914
Shelters	0.6261	0.5325	0.6180	0.5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0.3597
All GQs	0.7878	0.5057	0.9217	0.5153

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A telephone operation is in progress to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that were vacant during GQ Advanced Contact (GQAC) but were open on Census Day require imputation.

In addition, we will impute a pop size for GQs that have a reported Census Day population count that is much smaller than expected. Our initial proposal is to impute when the Census Day population count is 25% of the GQAC expected count, but research into determining this threshould (and refining it) is ongoing.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Visit but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with much lower than expected population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe.

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Additionally, some of the 185,000 resolved occupied GQs will be treated as unresolved because their census day population is much lower than expected. The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs as well as any GQs with a much lower than expected population count. Our current threshold for a "low" population count is < 25% of the GQAC expected count. Table 2 shows the distribution of the resolved and unresolved occupied GQS

Commented [JEZ(F1]: Tables based on 12/13/20 data.

Commented [PJC(F2]: For condition situation (2), do we have to mention other statuses?

Commented [JEZ(F3R2]: I think this question needs SME input. Are we accepting Vacant, Delete and Nonresidential GQ counts as-is?

Commented [PJC(F4]: We should explain, perhaps insert a note below the table or in the text that differentiates "Vacant During Visit, Open on Census Day" from "Vacant GQ"? I presume the two are mutually exclusive. So did the former report that they were open on CD, while the latter didn't report anything about CD? I don't understand.

Commented [JEZ(F5R4]: Need to ask Debbie and Ryan.

by GQ status. Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 9 in the Appendix has a full list of the GQ type codes.

Table 2: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Status	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	88,500	88,000	3,900	17,000	197,000
Vacant During Visit, Open on Census Day	1,000	550	300	19,500	21,500
Refusal GQ	350	450	300	6,700	7,800
Total	90,000	89,000	4,500	43,000	227,000

Commented [JEZ(F6]: Defined as pop count < 25% of expected. If the threshold is changed to < 10% of expected, count becomes 2,000.

Commented [JEZ(F7R6]: There also exist cases where the expected size is the same for all GQs in same facility. Sometimes these make sense, but sometimes it looks like they may be totals, when comparing to GP. For the unresolved, might not be able to tell.

Commented [JEZ(F8R6]: Might want to flag low count cases and do a manual review to determine if they may need imputation. Seems like expected count could have some measurement error issues, so we may not want to depend on it completely to determine if the CD pop is really too low.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Type	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	9,900	3,100	300	2,800	16,000
Juvenile Facilities	2,300	3,600	300	1,800	8,000
Nursing Facilities*	6,000	19,000	450	3,200	28,500
Hospitals	750	1,100	100	800	2,800
College Housing*	12,000	17,000	1,400	5,500	36,000
Military*	2,100	900	100	1,900	5,000
Shelters	21,000	3,200	550	8,200	33,000
Group Homes	29,000	32,500	850	9,100	72,000
Other	7,100	8,600	500	9,700	26,000
Total	90,000	89,000	4,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

An alternate or complimentary definition for a low census day population count would be to use 10% of the GQAC Max Number of People. Table 4 shows counts of the resolved and unresolved cases using this alternate threshold by GQ status. Table 5 shows the same information by GQ type.

Table 4: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Status	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	67,000	111,000	2,900	17,000	197,000
Vacant During Visit, Open on Census Day	550	1,000	350	19,500	21,500
Refusal GQ	150	650	300	6,700	7,800
Total	68,000	112,000	3,500	43,000	227,000

Commented [JEZ(F9]: 2,400 GQs have < 25% of expected count and < 10% of max count.

Table 5: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Type	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	5,600	7,200	400	2,800	16,000
Juvenile Facilities	1,600	4,400	150	1,800	8,000
Nursing Facilities*	4,300	20,500	300	3,200	28,500
Hospitals	550	1,300	90	800	2,800
College Housing*	7,800	21,500	1,200	5,500	36,000
Military*	1,500	1500	90	1,900	5,000
Shelters	17,000	7,300	300	8,200	33,000
Group Homes	24,000	38,500	450	9,100	72,000
Other	5,600	10,000	450	9,700	26,000
Total	68,000	112,000	3,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Commented [PJC(F10]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Variables

Table 6 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, and Administrative Records. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 6: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Vacant During Visit, Open on Census Day; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR

Additional sources available for college housing GQs include data collected via web-scraping data from the Integrated Postsecondary Education Data System (IPEDS) and data from the Common Core. These variables are available at the facility level but not for individual MAFIDs. For universities and colleges, we have the 2019 facility-level total room capacity (number of persons that could live in the GQ) from the IPEDS. To obtain these data….[information about CES matching]. The room capacity variable is of high-quality and is available for most (95%) of college housing.

Commented [JEZ(F11]: Is this only for 501s?

Table 7: Facility-Level Data available for College Housing GQs

Variable	Description	Source
[Web Scraping Vars]		
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Commented [JEZ(F12]: Need CES to provide details

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

First, if a pop count is available from the NPC call operation, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will use a combination of the following methods:

1. ~~Conversion from GQAC Expected Count~~Ratio Imputation
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

~~Conversion from GQAC~~Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will ~~substitute a function of those variables~~use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Commented [JEZ(F13]: Removed references to 'substitution'. When I wrote it originally, I meant we would use the exact value. I think others might have the same idea so I just removed that word and updated to say ratio imputation. I think that was today's decision – that we would adjust with a ratio and not use the values directly.

Table 8 shows that 8,600 of the unresolved cases can be resolved by substituting with the GQAC expected count.

For each GQ type, we will use the ratio of the reported GQ Census Day count to the GQAC expected count to convert the GQAC expected count of the unresolved GQ to a Census Day imputed count. For each GQ type, we will calculate the ratio of the sum of the GQAC Expected Count to the sum of the reported GQ population for the resolved cases. For the unresolved GQs, we will multiply the GQAC expected count by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People Current GQ Size and Max Number of People variables. We will not ~~substitute~~ use ratio imputation with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count.

Table 8: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Adjusted Residual from Facility-level Total for College Housing

If the GQ advance contact expected count is not populated, we will implement the following facility-level residual method. This method can only be used for GQs with GQTYPCUR=501 (colleges and universities) (For the rest of this section we use "college", "university", or "facility" to mean the same thing.

For 501s, we have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the 501 type GQs. The IPEDS room capacity may differ

Commented [TKW(F14]: Joe Staudt can provide a description of the matching algorithm, and the quality of the matches (which is very high for a high percentage of the cases).

from the college-level sum of GQ population counts for at least three reasons: (1) **reference year**—our latest IPEDS data is for reference year 2019; (2) **"capacity utilization"**—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day; (3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Adjusting the IPEDS facility-level Room Capacity

To adjust the IPEDS room capacity for reference year differences, we use the GQAC Max Number of People. We first select colleges for which we have a positive GQAC Max Number of People for every GQ at the facility. Since the IPEDS data does not include off-campus housing, we further subset on facilities that have no Greek letter GQs (fraternity or sorority houses). Finally, to maximize the chances that we are comparing apples to apples, we also subset to facilities for which the match quality is very high (match score > 90%). Within this subset, we calculate the average ratio of the facility-level sum of GQAC Max Number of People over the room capacity from IPEDS:

$$Average\ Ratio_S = \sum_{i \in S} \frac{\sum_{facility\_i} GQAC\ Max\ Number\ of\ People}{IPEDS\ Room\ Capacity\ at\ facility\ i}$$

where *S* is the set of facilities with no Greek GQs only positive values for GQAC Max Number of People.

Reassuringly, within this set of facilities, the median ratio is [redacted], the mode is [redacted], the 25th percentile is [redacted], and the 75th percentile is [redacted].

After adjusting the IPEDS college-level room capacity, we will similarly adjust for GQ "capacity utilization" at the college-level, using the mean ratio of 2020 Census Day GQ population over GQAC Max Number of People for all GQs for which both 2020 Census Day GQ population over GQAC Max Number of People. If time and sample sizes permit, we will also calculate this average ratio for college size classes. If the mean ratios differ significantly by college size class we will use separate capacity utilization adjustment for each college size class

After adjusting the college-level total room capacity to account reference year for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Finally, we will adjust the room capacity for GQ population in off-campus Greek housing (which is not included in the IPEDS room capacity). About 51% of colleges in the GQ data have no Greek letter GQs.

However, among colleges with at least 1 Greek letter GQ, at the mean has 38% of GQs are Greek letter houses, with a standard deviation of 34%. Since the importance of Greek letter GQs varies widely across colleges, we apply a Greek housing adjustment to each college based on which of 5 categories the colleges falls into:

1. No Greek housing GQs
2. Small school, low percentage of Greek housing GQs
3. Small school, high percentage of Greek housing GQs
4. Large school, low percentage of Greek housing GQs
5. Small school, high percentage of Greek housing GQs

For colleges with no/low GQ missingness rates, take average within each category of Greek housing pop counts over total GQ pop counts.

Imputing GQ-level population counts from the college-level residual

There are four possible cases:

1. For facililties with only one GQ with missing population, we will then impute the GQ population with:

$$Imputed\ Population\ Count = \ Residual_C$$

For facilities with more than one unresolved GQ, we will need to impute the fraction of the facility-level residual population that goes with each GQ. We propose a hierarchy of three approaches:

2-4. Need to fill these in with the descriptions from the top of my program 07.*.sas on IRE /projects/GQ_Imputation/Kirk/

Commented [JEZ(F15]: From Kirk.

Modeling

If ~~the previous two steps do not yield an imputation (no GQAC expected count and no IPEDS count) for the unresolved GQ and~~ sufficient auxiliary variables are available, we will impute with a prediction from a logistic or Poisson regression model. For the logistic regression the dependent variable will be the reported count / max number of people (because it is the most often filled size variable). For the Poisson regression, the dependent variable will be reported GQ pop count with an offset of the max number of people. Independent variables will be selected from Table 6. It is important to note that GQ type will either be a covariate in the models or separate models will be fit by GQ type. Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

Median Imputation

If sufficient auxiliary data is not available, we will impute the pop size with median population within an imputation cell. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and GQ status. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

Commented [JEZ(F16]: Need to add text about comparing all methods and determining a hierarchy.

Models

We will evaluate the imputation models using cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The remaining 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will predict the GQ pop size for the validation deck. For the validation deck, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each GQ. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

~~Conversion~~ Ratio Imputation and Median

To evaluate the ~~Conversion from GQAC~~Ratio Imputation and median methods, we will simply conduct the procedures and then review them for reasonableness.

Appendices

Table 9: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0 3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1.017	0.6385
College Housing	0.7818	0.5492	0.9444	0 5535
Military	0.7317	0.2290	0.9492	0 2914
Shelters	0.6261	0.5325	0.6180	0 5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0 3597
All GQs	0.7878	0.5057	0.9217	0 5153

Commented [JEZ(F17]: Probably need to set thresholds and remove outliers before determining ratios to use for GQAC conversion.

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

Background

Commented [JEZ(F1]: Tables based on 12/13/20 data.

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A telephone operation is in progress to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that were vacant during GQ Advanced Contact (GQAC) but were open on Census Day require imputation.

In addition, we will impute a pop size for GQs that have a reported Census Day population count that is much smaller than expected. Our initial proposal is to impute when the Census Day population count is 25% of the GQAC expected count, but research into determining this threshould (and refining it) is ongoing.

Imputation Universe

Commented [JEZ(F2]: Meeting comments: Need to decide if we are imputing for low counts or only missing/zero counts. If we are going to impute for discrepancy cases, what are the conditions or thresholds?

Need to address which cases we will use treat as resolved for imputation – i.e. “donors”.

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Visit but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with much lower than expected population count are included in the Census Day Pop column.

Commented [PJC(F3]: For condition situation (2), do we have to mention other statuses?

Commented [JEZ(F4R3]: I think this question needs SME input. Are we accepting Vacant, Delete and Nonresidential GQ counts as-is?

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Commented [PJC(F5]: We should explain, perhaps insert a note below the table or in the text that differentiates "Vacant During Visit, Open on Census Day" from "Vacant GQ"? I presume the two are mutually exclusive. So did the former report that they were open on CD, while the latter didn't report anything about CD? I don't understand.

Commented [JEZ(F6R5]: Need to ask Debbie and Ryan.

Table 2 shows the status of the occupied GQs. There are 43,000 unresolved occupied GQs without a census day population.

Table 2: Occupied GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Total	184,000	43,000	227,000

Additionally some of the 184,000 resolved occupied GQs will be treated as unresolved because their census day population is much lower than expected.

Commented [TLK(F7]: We need to calculate how many GQs have a Census Day population that is 25% of the GQAC expected count.

Commented [PJC(F8]: Good work. Great table, just below. After review, we can decide if 25% is an appropriate threshold for Low Census Day Pop.

GQ Status	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	88,500	88,000	3,900	17,000	197,000
Vacant During Visit, Open on Census Day	1,000	550	300	19,500	21,500
Refusal GQ	350	450	300	6,700	7,800
Total	90,000	89,000	4,500	43,000	227,000

Commented [JEZ(F9]: Defined as pop count < 25% of expected. If the threshold is changed to < 10% of expected, count becomes 2,000.

Commented [JEZ(F10R9]: There also exist cases where the expected size is the same for all GQs in same facility. Sometimes these make sense, but sometimes it looks like they may be totals, when comparing to GP. For the unresolved, might not be able to tell.

Commented [JEZ(F11R9]: Might want to flag low count cases and do a manual review to determine if they may need imputation. Seems like expected count could have some measurement error issues, so we may not want to depend on it completely to determine if the CD pop is really too low.

Commented [JEZ(F12]: 100 of these have expected size <= 5. An additional 350 have expected size between 6 and 10.

GQ Type	Low Census Day Pop
Correctional Facilities	300
Juvenile Facilities	300
Nursing Facilities	450
Hospitals	100
College Housing	1,400
Military	100
Shelters	550
Group Homes	850
Other	500
All GQs	4,500

The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs as well as any GQs with a much lower than expected population count. Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 7 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

GQ Type	Census Day Pop	Unresolved	Total
Correctional Facilities*	13,000	2,800	16,000
Juvenile Facilities	6,200	1,800	8,000
Nursing Facilities*	25,500	3,200	28,500
Hospitals	2,000	800	2,800
College Housing*	30,500	5,500	36,000
Military*	3,100	1,900	5,000
Shelters	24,500	8,200	33,000
Group Homes	62,500	9,100	72,000

Commented [PJC(F13]: So these frequencies will be updated to include GQs with unresolved Low Census Day Pop, correct?

Other	16,000	9,700	26,000
Total	184,000	43,000	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Commented [PJC(F14]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Variables

Table 4 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, and Administrative Records. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 4: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQ Advanced Contact	Master Address File / DRF1
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Vacant During Visit, Open on Census Day; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR

Commented [PJC(F15]: How does this differ from the prior variable? Also, what is the name of this variable as collected from other sources, such as IPEDS, if IPEDS or the others sources have the information at a GQ level?

Additional sources available for college housing GQs include data collected via web-scraping, data from the Integrated Postsecondary Education Data System (IPEDS) and data from the Common Core. These variables are available at the facility level but not for individual MAFIDs. For universities and colleges, we have the 2019 facility-level total room capacity (number of persons that could live in the GQ) from the IPEDS. To obtain these data….[information about CES matching]. The room capacity variable is of high-quality and is available for most (95%) of college housing.

Commented [JEZ(F16]: Is this only for 501s?

Table 5: Facility-Level Data available for College Housing GQs

Variable	Description	Source
[Web Scraping Vars]		

Commented [JEZ(F17]: Need CES to provide details

Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

First, if a pop count is available from the NPC call operation, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will be hierarchical, following three steps:

1. Conversion from GQAC Expected Count
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Mean Imputation

Conversion from GQAC

For cases where we have an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will substitute a function of those variables. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 6 shows that 8,600 of the unresolved cases can be resolved by substituting with the GQAC expected count.

For each GQ type, we will use the ratio of the reported GQ Census Day count to the GQAC expected count to convert the GQAC expected count of the unresolved GQ to a Census Day imputed count. For each GQ type, we will calculate the ratio of the sum of the GQAC Expected Count to the sum of the reported GQ population for the resolved cases. For the unresolved GQs, we will multiply the GQAC expected count by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will not substitute with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count.

Table 6: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Adjusted Residual from Facility-level Total for College Housing

If the GQ advance contact expected count is not populated, we will implement the following facility-level residual method. This method can only be used for GQs with GQTYPCUR=501.

For universities and colleges, we have the 2019 facility-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the facility level to the GQ data. We will adjust the room capacity for GQ population in off-campus Greek housing (which is not included in the IPEDS room capacity). To avoid overcounting, we will also scale the room capacity by the average ratio (within facility size classes) of the facility-level total 2020 Census Day population over the facility-level room capacity. For calculating these ratios, we will only use facilities with for which less than 5 or 10% of the GQs at the facility are unresolved cases.

Commented [TKW(F18]: Joe Staudt can provide a description of the matching algorithm, and the quality of the matches (which is very high for a high percentage of the cases).

How to portion out the residual to multiple GQs

For facilities with more than one unresolved GQ, we will need to impute the fraction of the facility-level residual population that goes with each GQ. We propose a hierarchy of two approaches:

1. We will sum the reported GQ population counts from the 2010 Decennial to facility level. (This data has already been merged on mafid to the 2020 GQ counts file. Then for each GQ we will calculate its share of the facility's population. For GQs (mafid) that existed in 2010 and still exist in 2020 we will these 2010 GQ shares of facility-level population to calculate the share of the facility's residual population (calculated as described above) at each unresolved GQ. For any unresolved GQs that cannot be imputed this way, we will follow approach 2.
2. For unresolved GQs that did not exist in 2010 (and for which we have no 2020 GQ-level estimate), we will divide the residual facility-level population evenly among the remaining GQs.

Modeling

If the previous two steps do not yield an imputation (no GQAC expected count and no IPEDS count) for the unresolved GQ and sufficient auxiliary variables are available, we will impute with a prediction from a logistic or Poisson regression model. For the logistic regression the dependent variable will be the reported count / max number of people (because it is the most often filled size variable). For the Poisson regression, the dependent variable will be reported GQ pop count with an offset of the max number of people. Independent variables will be selected from Table 4. It is important to note that GQ type will either be a covariate in the models or separate models will be fit by GQ type. Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

Mean Imputation

If sufficient auxiliary data is not available, we will impute the pop size with average population within an imputation cell. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and GQ status. Then, we will calculate the average GQ population size and impute the unresolved GQs with the average.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

Models

We will evaluate the imputation models using cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The remaining 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will predict the GQ pop size for the validation deck. For the validation deck, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each GQ. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Conversion and Mean

To evaluate the Conversion from GQAC and mean methods, we will simply conduct the procedures and then review them for reasonableness.

Appendices

Table 7: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0 3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1 017	0.6385
College Housing	0.7818	0.5492	0.9444	0.5535
Military	0.7317	0.2290	0.9492	0.2914
Shelters	0.6261	0.5325	0.6180	0.5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0.3597
All GQs	0.7878	0.5057	0.9217	0.5153

Commented [JEZ(F19]: Probably need to set thresholds and remove outliers before determining ratios to use for GQAC conversion.

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Background

Commented [JEZ(F1]: Tables based on 12/13/20 data.

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A telephone operation is in progress to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that were vacant during GQ Advanced Contact (GQAC) but were open on Census Day require imputation.

In addition, we will impute a pop size for GQs that have a reported Census Day population count that is much smaller than expected. Our initial proposal is to impute when the Census Day population count is 25% of the GQAC expected count, but research into determining this threshould (and refining it) is ongoing.

Imputation Universe

Commented [JEZ(F2]: Meeting comments: Need to decide if we are imputing for low counts or only missing/zero counts. If we are going to impute for discrepancy cases, what are the conditions or thresholds?

Need to address which cases we will use treat as resolved for imputation – i.e. "donors".

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Visit but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with much lower than expected population count are included in the Census Day Pop column.

Commented [PJC(F3]: For condition situation (2), do we have to mention other statuses?

Commented [JEZ(F4R3]: I think this question needs SME input. Are we accepting Vacant, Delete and Nonresidential GQ counts as-is?

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Commented [PJC(F5]: We should explain, perhaps insert a note below the table or in the text that differentiates "Vacant During Visit, Open on Census Day" from "Vacant GQ"? I presume the two are mutually exclusive. So did the former report that they were open on CD, while the latter didn't report anything about CD? I don't understand.

Commented [JEZ(F6R5]: Need to ask Debbie and Ryan.

Table 2 shows the status of the occupied GQs. There are 43,000 unresolved occupied GQs without a census day population.

Table 2: Occupied GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000

Vacant During Visit, Open on Census Day	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Total	184,000	43,000	227,000

Additionally some of the 184,000 resolved occupied GQs will be treated as unresolved because their census day population is much lower than expected.

Commented [TLK(F7]: We need to calculate how many GQs have a Census Day population that is 25% of the GQAC expected count.

GQ Status	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	88,500	88,000	3,900	17,000	197,000
Vacant During Visit, Open on Census Day	1,000	550	300	19,500	21,500
Refusal GQ	350	450	300	6,700	7,800
Total	90,000	89,000	4,500	43,000	227,000

Commented [JEZ(F8]: Defined as pop count < 25% of expected. If the threshold is changed to < 10% of expected, count becomes 2,000.

Commented [JEZ(F9R8]: There also exist cases where the expected size is the same for all GQs in same facility. Sometimes these make sense, but sometimes it looks like they may be totals, when comparing to GP. For the unresolved, might not be able to tell.

Commented [JEZ(F10R8]: Might want to flag low count cases and do a manual review to determine if they may need imputation. Seems like expected count could have some measurement error issues, so we may not want to depend on it completely to determine if the CD pop is really too low.

Commented [JEZ(F11]: 100 of these have expected size <= 5. An additional 350 have expected size between 6 and 10.

GQ Type	Low Census Day Pop
Correctional Facilities	300
Juvenile Facilities	300
Nursing Facilities	450
Hospitals	100
College Housing	1,400
Military	100
Shelters	550
Group Homes	850
Other	500
All GQs	4,500

The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs as well as any GQs with a much lower than expected population count. Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 7 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

GQ Type	Census Day Pop	Unresolved	Total
Correctional Facilities*	13,000	2,800	16,000
Juvenile Facilities	6,200	1,800	8,000
Nursing Facilities*	25,500	3,200	28,500
Hospitals	2,000	800	2,800
College Housing*	30,500	5,500	36,000
Military*	3,100	1,900	5,000
Shelters	24,500	8,200	33,000
Group Homes	62,500	9,100	72,000
Other	16,000	9,700	26,000

Total	184,000	43,000	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Variables

Table 4 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, and Administrative Records. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 4: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQ Advanced Contact	Master Address File / DRF1
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Vacant During Visit, Open on Census Day; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR

Additional sources available for college housing GQs include data collected via web-scraping, data from the Integrated Postsecondary Education Data System (IPEDS) and data from the Common Core. These variables are available at the facility level but not for individual MAFIDs. For universities and colleges, we have the 2019 facility-level total room capacity (number of persons that could live in the GQ) from the IPEDS. To obtain these data....[information about CES matching]. The room capacity variable is of high-quality and is available for most (95%) of college housing.

Commented [JEZ(F12]: Is this only for 501s?

Table 5: Facility-Level Data available for College Housing GQs

Variable	Description	Source
[Web Scraping Vars]		
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for,	IPEDS

Commented [JEZ(F13]: Need CES to provide details

	whether on or off campus (off-campus dormitory space that is reserved by the institution).	

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

First, if a pop count is available from the NPC call operation, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will be hierarchical, following three steps:

1. Conversion from GQAC Expected Count
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Mean Imputation

Conversion from GQAC

For cases where we have an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will substitute a function of those variables. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 6 shows that 8,600 of the unresolved cases can be resolved by substituting with the GQAC expected count.

For each GQ type, we will use the ratio of the reported GQ Census Day count to the GQAC expected count to convert the GQAC expected count of the unresolved GQ to a Census Day imputed count. For each GQ type, we will calculate the ratio of the sum of the GQAC Expected Count to the sum of the reported GQ population for the resolved cases. For the unresolved GQs, we will multiply the GQAC expected count by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will not substitute with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count.

Table 6: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Adjusted Residual from Facility-level Total for College Housing

Formatted: Normal

If the GQ advance contact expected count is not populated, we will implement the following facility-level residual method. This method can only be used for GQs with GQTYPCUR=501 (colleges and universities)~~.~~ (For the rest of this section we use "college" "university" or "facility" to mean the same thing.

For ~~universities and colleges~~501s, we have the 2019 ~~facility~~college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the ~~facility~~ college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the 501 type GQs. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons: (1) **reference year**—our latest IPEDS data is for reference year 2019; (2) "**capacity utilization**"—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day; (3) **scope**—IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Commented [TKW(F14]: Joe Staudt can provide a description of the matching algorithm, and the quality of the matches (which is very high for a high percentage of the cases).

Formatted: Font: Bold

Formatted: Font: Bold

Formatted: Font: Italic

Formatted: Font: Bold

Adjusting the IPEDS facility-level Room Capacity

To adjust the IPEDS room capacity for reference year differences we use the GQAC Max Number of People. We first select colleges for which we have a positive GQAC Max Number of People for every GQ at the facility. Since the IPEDS data does not include off-campus housing, we further subset on facilities that have no Greek letter GQs (fraternity or sorority houses). Finally, to maximize the chances that we are comparing apples to apples we also subset to facilities for which the match quality is very high (match score > 90%). Within this subset we calculate the average ratio of the facility-level sum of GQAC Max Number of People over the room capacity from IPEDS:

Formatted: Font: 11 pt

Formatted: Font: 11 pt

Formatted: Font: 11 pt

$$Average\ Ratio_S = \sum_{i \in S} \frac{\sum_{facility\_i} GQAC\ Max\ Number\ of\ People}{IPEDS\ Room\ Capacity\ at\ facility\ i}$$

where *S* is the set of facilities with no Greek GQs only positive values for GQAC Max Number of People.

Formatted: Font: 11 pt

Formatted: Font: 11 pt

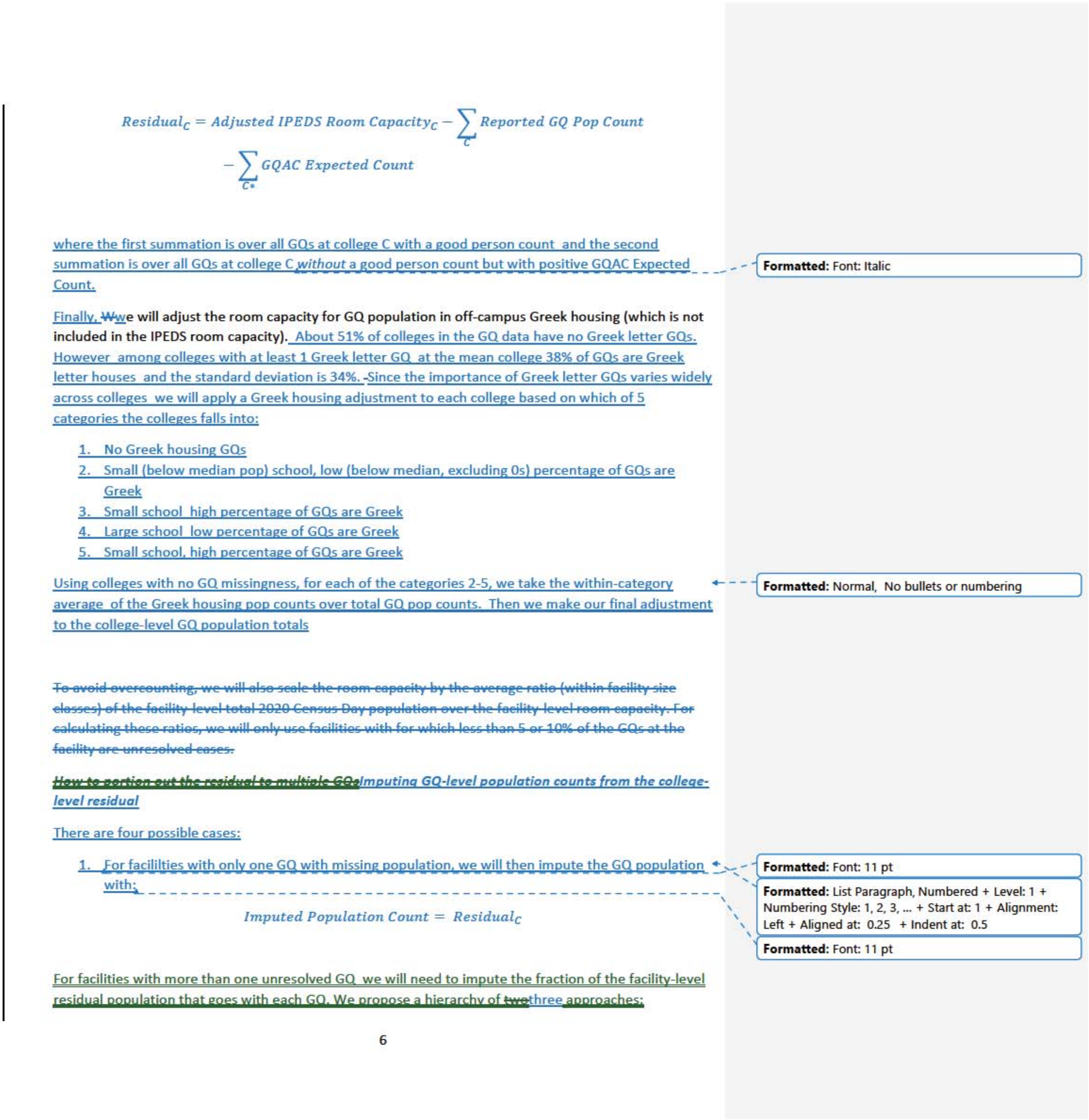
Reassuringly, within this set of facilities, the median ratio is [redacted], the mode is [redacted], the 25th percentile is [redacted] and the 75th percentile is [redacted].

After adjusting the IPEDS college-level room capacity we will similarly adjust for GQ "capacity utilization" at the college-level using the mean ratio of 2020 Census Day GQ population over GQAC Max Number of People for all GQs for which both 2020 Census Day GQ population over GQAC Max Number of People. If time and sample sizes permit, we will also calculate this average ratio for college size classes. If the mean ratios differ significantly by college size class we will use separate capacity utilization adjustment for each college size class

Formatted: Font: 11 pt

Formatted: Font: 11 pt

After adjusting the college-level total room capacity to account reference year for capacity utilization we will calculate the following college-level residual for each college C:



$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Formatted: Font: Italic

Finally, ~~W~~we will adjust the room capacity for GQ population in off-campus Greek housing (which is not included in the IPEDS room capacity). About 51% of colleges in the GQ data have no Greek letter GQs. However among colleges with at least 1 Greek letter GQ at the mean college 38% of GQs are Greek letter houses and the standard deviation is 34%. Since the importance of Greek letter GQs varies widely across colleges we will apply a Greek housing adjustment to each college based on which of 5 categories the colleges falls into:

1. No Greek housing GQs
2. Small (below median pop) school, low (below median, excluding 0s) percentage of GQs are Greek
3. Small school high percentage of GQs are Greek
4. Large school low percentage of GQs are Greek
5. Small school, high percentage of GQs are Greek

Using colleges with no GQ missingness, for each of the categories 2-5, we take the within-category average of the Greek housing pop counts over total GQ pop counts. Then we make our final adjustment to the college-level GQ population totals

Formatted: Normal, No bullets or numbering

~~To avoid overcounting, we will also scale the room capacity by the average ratio (within facility size classes) of the facility-level total 2020 Census Day population over the facility-level room capacity. For calculating these ratios, we will only use facilities with for which less than 5 or 10% of the GQs at the facility are unresolved cases.~~

~~How to portion out the residual to multiple GQs~~Imputing GQ-level population counts from the college-level residual

There are four possible cases:

1. For facililties with only one GQ with missing population, we will then impute the GQ population with;

$$Imputed\ Population\ Count = Residual_C$$

Formatted: Font: 11 pt

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.25 + Indent at: 0.5

Formatted: Font: 11 pt

For facilities with more than one unresolved GQ we will need to impute the fraction of the facility-level residual population that goes with each GQ. We propose a hierarchy of ~~two~~three approaches:

2-4. Need to fill these in with the descriptions from the top of my program 07.*.sas on IRE /projects/GQ_Imputation/Kirk/

~~***How to portion out the residual to multiple GQs***~~

~~For facilities with more than one unresolved GQ, we will need to impute the fraction of the facility-level residual population that goes with each GQ. We propose a hierarchy of two approaches:~~

1. ~~We will sum the reported GQ population counts from the 2010 Decennial to facility level. (This data has already been merged on mafid to the 2020 GQ counts file. Then for each GQ we will calculate its share of the facility's population. For GQs (mafid) that existed in 2010 and still exist in 2020 we will these 2010 GQ shares of facility level population to calculate the share of the facility's residual population (calculated as described above) at each unresolved GQ. For any unresolved GQs that cannot be imputed this way, we will follow approach 2.~~
2. ~~For unresolved GQs that did not exist in 2010 (and for which we have no 2020 GQ-level estimate), we will divide the residual facility-level population evenly among the remaining GQs.~~

Modeling

If the previous two steps do not yield an imputation (no GQAC expected count and no IPEDS count) for the unresolved GQ and sufficient auxiliary variables are available, we will impute with a prediction from a logistic or Poisson regression model. For the logistic regression the dependent variable will be the reported count / max number of people (because it is the most often filled size variable). For the Poisson regression, the dependent variable will be reported GQ pop count with an offset of the max number of people. Independent variables will be selected from Table 4. It is important to note that GQ type will either be a covariate in the models or separate models will be fit by GQ type. Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

Mean Imputation

If sufficient auxiliary data is not available, we will impute the pop size with average population within an imputation cell. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and GQ status. Then, we will calculate the average GQ population size and impute the unresolved GQs with the average.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

Models

We will evaluate the imputation models using cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The remaining 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will predict the GQ pop size for the validation deck. For the validation deck, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each GQ. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Conversion and Mean

To evaluate the Conversion from GQAC and mean methods, we will simply conduct the procedures and then review them for reasonableness.

Appendices

Table 7: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0 3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1.017	0.6385
College Housing	0.7818	0.5492	0.9444	0 5535
Military	0.7317	0.2290	0.9492	0 2914
Shelters	0.6261	0.5325	0.6180	0 5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0 3597
All GQs	0.7878	0.5057	0.9217	0 5153

Commented [JEZ(F15]: Probably need to set thresholds and remove outliers before determining ratios to use for GQAC conversion.

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A telephone operation is in progress to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that open on Census Day, but vacant during the GQ Enumeration visit (which started in July 2020) require imputation.

In addition, we will impute a pop size for GQs that have a reported Census Day population count that is much smaller than expected. Our initial proposal is to impute when the Census Day population count is 25% of the GQAC expected count, but research into determining (and refining) this threshould is ongoing.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Visit but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with much lower than expected population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe.

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Open on Census Day, Vacant During Visit	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Additionally, some of the 185,000 resolved occupied GQs will be treated as unresolved because their census day population is much lower than expected. The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs as well as any GQs with a much lower than expected population count. Our current threshold for a “low” population count is < 25% of the GQAC expected count. Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ status. Of the resolved GQs, 89,000 had a GQAC expected count and 90,000 did not. The

unresolved GQs include the 43,000 GQs without a reported count as well as 4,500 that had a large discrepancy between the GQAC expected population and the reported pop size.

Table 2: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Status	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	88,500	88,000	3,900	17,000	197,000
Vacant During Visit, Open on Census Day	1,000	550	300	19,500	21,500
Refusal GQ	350	450	300	6,700	7,800
Total	90,000	89,000	4,500	43,000	227,000

Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 10 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Type	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	9,900	3,100	300	2,800	16,000
Juvenile Facilities	2,300	3,600	300	1,800	8,000
Nursing Facilities*	6,000	19,000	450	3,200	28,500
Hospitals	750	1,100	100	800	2,800
College Housing*	12,000	17,000	1,400	5,500	36,000
Military*	2,100	900	100	1,900	5,000
Shelters	21,000	3,200	550	8,200	33,000
Group Homes	29,000	32,500	850	9,100	72,000
Other	7,100	8,600	500	9,700	26,000
Total	90,000	89,000	4,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

An alternate definition for a low census day population count would be to use 10% of the GQAC Max Number of People. **Error! Not a valid bookmark self-reference.** shows counts of the resolved and unresolved cases using this alternate threshold by GQ status. Table 5 shows the same information by GQ type. We will examine using the intersection or union of these conditions as well as setting thresholds at different levels to determine which reported counts require imputation.

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Table 4: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Status	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	67,000	111,000	2,900	17,000	197,000
Vacant During Visit, Open on Census Day	550	1,000	350	19,500	21,500
Refusal GQ	150	650	300	6,700	7,800
Total	68,000	112,000	3,500	43,000	227,000

Note that 2,400 GQs with the Low Census Day Pop based on the Max Pop also have a Low Census Day Pop using the GQAC Expected Population.

Table 5: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Type	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	5,600	7,200	400	2,800	16,000
Juvenile Facilities	1,600	4,400	150	1,800	8,000
Nursing Facilities*	4,300	20,500	300	3,200	28,500
Hospitals	550	1,300	90	800	2,800
College Housing*	7,800	21,500	1,200	5,500	36,000
Military*	1,500	1500	90	1,900	5,000
Shelters	17,000	7,300	300	8,200	33,000
Group Homes	24,000	38,500	450	9,100	72,000
Other	5,600	10,000	450	9,700	26,000
Total	68,000	112,000	3,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Variables

Table 6 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, and Administrative Records. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 6: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR

Additional sources available for college housing GQs include data collected via web-scraping, data from the Integrated Postsecondary Education Data System (IPEDS) and data from the Common Core. These variables are available at the facility level but not for individual MAFIDs.

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) "**capacity utilization**"—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Additional facility-level variables may become available as research continues.

Table 7: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

First, if a pop count is available from the NPC call operation, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will use a combination of the following methods:

1. Ratio Imputation
2. Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 8 shows that 8,600 of the unresolved GQ can be resolved by converting the GQAC expected count to the GQ pop count using the following ratio adjustment.

Table 8: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For each GQ type, we will use the ratio of the reported GQ Census Day count to the GQAC expected count to convert the GQAC expected count of the unresolved GQ to a Census Day imputed count. For each GQ type, we will calculate the ratio of the sum of the GQAC Expected Count to the sum of the reported GQ population for the resolved cases. For the unresolved GQs, we will multiply the GQAC expected count by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. We will not use ratio imputation with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count. Conversion factors for the four variables under consideration are

shown in Table 9. Table 12Table 14 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 9: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0.3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1.017	0.6385
College Housing	0.7818	0.5492	0.9444	0.5535
Military	0.7317	0.2290	0.9492	0.2914
Shelters	0.6261	0.5325	0.6180	0.5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0.3597
All GQs	0.7878	0.5057	0.9217	0.5153

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for GQs for colleges and universities (GQTYPCUR=501).

First, we will adjust the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population, GQAC Max Number of People, and Greek Housing variables.

After adjusting the college-level total room capacity to account reference year and for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Once we calculate the college-level residual, we will then allocate the population counts among the GQs in the college without GQAC Expected Count.

Modeling

A third approach would be to impute the GQ pop counts from a Poisson regression model. The dependent variable will be reported GQ pop count with an offset of the max number of people (because that is filled the most). Independent variables will be selected from Table 6. It is important to note that GQ type will either be a fixed-effect covariate in the models or separate models will be fit by GQ type.

Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

Median Imputation

If sufficient auxiliary data is not available, we will impute the pop size with median population within an imputation cell. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and GQ status. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median GQ pop size in the cell.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

We will evaluate the imputation methods using cross validation. First, we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The remaining 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will impute the GQ pop size for all GQs in the validation deck. That is, we will attempt to impute the GQ pop size for every GQ in the 60% sample four times (once for each of the four methods). Note that the second method can only be applied to college housing. Then, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each method. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Some methods may perform better than others for certain types of units. For example, Poisson regression might perform best when the GQAC expected count is available, but not well when it is missing. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Appendix

Table 10: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Table 11: GQAC Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Table 12: GQAC Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Table 13: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Table 14: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

Background

Commented [JEZ(F1]: Tables based on 12/13/20 data.

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A telephone operation is in progress to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that open on Census Day, but vacant during the GQ Enumeration visit (which started in July 2020) require imputation.

In addition, we will impute a pop size for GQs that have a reported Census Day population count that is much smaller than expected. Our initial proposal is to impute when the Census Day population count is 25% of the GQAC expected count, but research into determining (and refining) this threshould is ongoing.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Visit but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with much lower than expected population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe.

Commented [PJC(F2]: For condition situation (2), do we have to mention other statuses?

Commented [JEZ(F3R2]: I think this question needs SME input. Are we accepting Vacant, Delete and Nonresidential GQ counts as-is?

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Open on Census Day, Vacant During Visit	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Commented [PJC(F4]: We should explain, perhaps insert a note below the table or in the text that differentiates "Vacant During Visit, Open on Census Day" from "Vacant GQ"? I presume the two are mutually exclusive. So did the former report that they were open on CD, while the latter didn't report anything about CD? I don't understand.

Commented [JEZ(F5R4]: Need to ask Debbie and Ryan.

Additionally, some of the 185,000 resolved occupied GQs will be treated as unresolved because their census day population is much lower than expected. The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs as well as any GQs with a much lower than expected population count. Our current threshold for a "low" population count is < 25% of the GQAC expected count. Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ status. Of the resolved GQs, 89,000 had a GQAC expected count and 90,000 did not. The

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

unresolved GQs include the 43,000 GQs without a reported count as well as 4,500 that had a large discrepancy between the GQAC expected population and the reported pop size.

Table 2: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Status	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	88,500	88,000	3,900	17,000	197,000
Vacant During Visit, Open on Census Day	1,000	550	300	19,500	21,500
Refusal GQ	350	450	300	6,700	7,800
Total	90,000	89,000	4,500	43,000	227,000

Commented [JEZ(F6]: Defined as pop count < 25% of expected. If the threshold is changed to < 10% of expected, count becomes 2,000.

Commented [JEZ(F7R6]: There also exist cases where the expected size is the same for all GQs in same facility. Sometimes these make sense, but sometimes it looks like they may be totals, when comparing to GP. For the unresolved, might not be able to tell.

Commented [JEZ(F8R6]: Might want to flag low count cases and do a manual review to determine if they may need imputation. Seems like expected count could have some measurement error issues, so we may not want to depend on it completely to determine if the CD pop is really too low.

Commented [JEZ(F9]: 100 of these have expected size <= 5. An additional 350 have expected size between 6 and 10.

Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 10 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Type	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	9,900	3,100	300	2,800	16,000
Juvenile Facilities	2,300	3,600	300	1,800	8,000
Nursing Facilities*	6,000	19,000	450	3,200	28,500
Hospitals	750	1,100	100	800	2,800
College Housing*	12,000	17,000	1,400	5,500	36,000
Military*	2,100	900	100	1,900	5,000
Shelters	21,000	3,200	550	8,200	33,000
Group Homes	29,000	32,500	850	9,100	72,000
Other	7,100	8,600	500	9,700	26,000
Total	90,000	89,000	4,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

An alternate or complimentary definition for a low census day population count would be to use 10% of the GQAC Max Number of People. Applying this definition would result in 1,100 more unresolved GQs, in addition to the 43,000 and 4,500 unresolved GQs in Table 3. Table 4 shows counts of the resolved and unresolved cases using this alternate threshold by GQ status. Table 5 shows the same information by GQ type.

Table 4: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Status	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	67,000	111,000	2,900	17,000	197,000
Vacant During Visit, Open on Census Day	550	1,000	350	19,500	21,500
Refusal GQ	150	650	300	6,700	7,800
Total	68,000	112,000	3,500	43,000	227,000

Note that 2,400 GQs with the Low Census Day Pop based on the Max Pop also have a Low Census Day Pop using the GQAC Expected Population.

Commented [JEZ(F10]: 2,400 GQs have < 25% of expected count and < 10% of max count.

Table 5: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Type	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	5,600	7,200	400	2,800	16,000
Juvenile Facilities	1,600	4,400	150	1,800	8,000
Nursing Facilities*	4,300	20,500	300	3,200	28,500
Hospitals	550	1,300	90	800	2,800
College Housing*	7,800	21,500	1,200	5,500	36,000
Military*	1,500	1500	90	1,900	5,000
Shelters	17,000	7,300	300	8,200	33,000
Group Homes	24,000	38,500	450	9,100	72,000
Other	5,600	10,000	450	9,700	26,000
Total	68,000	112,000	3,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Commented [PJC(F11]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Variables

Table 6 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, and Administrative Records. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 6: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR

Additional sources available for college housing GQs include data collected via web-scraping, data from the Integrated Postsecondary Education Data System (IPEDS) and data from the Common Core. These variables are available at the facility level but not for individual MAFIDs.

Commented [JEZ(F12]: Is this only for 501s?

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) “**capacity utilization**”—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**—IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Additional facility-level variables may become available as research continues.

Table 7: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

First, if a pop count is available from the NPC call operation, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will use a combination of the following methods:

1. Ratio Imputation
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

Commented [ADK(F13]: Need to look at paradata as covariates as well on the models

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 8 shows that 8,600 of the unresolved GQ can be resolved by converting the GQAC expected count to the GQ pop count using the following ratio adjustment.

Table 8: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For each GQ type, we will use the ratio of the reported GQ Census Day count to the GQAC expected count to convert the GQAC expected count of the unresolved GQ to a Census Day imputed count. For each GQ type, we will calculate the ratio of the sum of the GQAC Expected Count to the sum of the reported GQ population for the resolved cases. For the unresolved GQs, we will multiply the GQAC expected count by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. We will not use ratio imputation with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count. Conversion factors for the four variables under consideration are

shown in Table 9. Tables 12-14 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 9: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0.3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1.017	0.6385
College Housing	0.7818	0.5492	0.9444	0.5535
Military	0.7317	0.2290	0.9492	0.2914
Shelters	0.6261	0.5325	0.6180	0.5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0.3597
All GQs	0.7878	0.5057	0.9217	0.5153

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for GQs for colleges and universities (GQTYPCUR=501).

Adjusting the IPEDS facility-level Room Capacity

To adjust the IPEDS room capacity for reference year differences, we use the GQAC Max Number of People. We first select colleges for which we have a positive GQAC Max Number of People for every GQ at the facility. Since the IPEDS data does not include off-campus housing, we further subset on facilities that have no Greek letter GQs (fraternity or sorority houses). Finally, to maximize the chances that we are comparing apples to apples, we also subset to facilities for which the match quality is very high (match score > 90%). Within this subset, we calculate the average ratio of the facility-level sum of GQAC Max Number of People over the room capacity from IPEDS:

$$Average\ Ratio_S = \sum_{i \in S} \frac{\sum_{college_i} GQAC\ Max\ Number\ of\ People}{IPEDS\ Room\ Capacity\ at\ college\ i}$$

where S is the set of colleges with no Greek GQs only positive values for GQAC Max Number of People.

Reassuringly, within this set of colleges, the median ratio is [redacted], the mode is [redacted], the 25th percentile is [redacted], and the 75th percentile is [redacted].

After adjusting the IPEDS college-level room capacity, we will similarly adjust for GQ "capacity utilization" at the college-level, using the mean ratio of 2020 Census Day GQ population over GQAC Max Number of People ~~for all~~among ~~GQs~~ colleges for which ~~both~~all 2020 Census Day GQ populations ~~over~~ and GQAC Max Number of People are non-missing and positive. If time and sample sizes permit, we will also calculate this average ratio for college size classes. If the mean ratios differ significantly by college size class we will use separate capacity utilization adjustment for each college size class

Commented [TLK(F14]: I think a word is missing from this sentence.

Commented [TKW(F15]: I have added the missing words now.

After adjusting the college-level total room capacity to account reference year for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Finally, we will adjust the room capacity for GQ population in off-campus Greek housing (which is not included in the IPEDS room capacity). About 51% of colleges in the GQ data have no Greek letter GQs. However, among colleges with at least 1 Greek letter GQ, ~~at the mean has~~the mean of the distribution of Greek letter GQs as a percentage of all GQs at the college is 38% ~~of GQs are Greek letter houses, with a~~and the standard deviation ~~of~~is 34%. Since the importance of Greek letter GQs varies widely across colleges, we apply a Greek housing adjustment to each college based on which of 5 categories the colleges falls into:

Commented [TLK(F16]: Clarify this.

1. No Greek housing GQs
2. Small school, low percentage of Greek housing GQs
3. Small school, high percentage of Greek housing GQs
4. Large school, low percentage of Greek housing GQs
5. Small school, high percentage of Greek housing GQs

Using colleges with no GQ missingness, for each of the categories 2-5, we take the within-category average of the Greek housing pop counts over total GQ pop counts. Then we will make our final adjustment to the college-level GQ population totals.

Once we calculate the collete-level residual, we will then allocate the population counts among the GQs in the college.

Modeling

A third approach would be to impute the GQ pop counts from a Poisson regression model. The dependent variable will be reported GQ pop count with an offset of the max number of people (because that is filled the most). Independent variables will be selected from Table 6. It is important to note that GQ type will either be a fixed-effect covariate in the models or separate models will be fit by GQ type. Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

Commented [ADK(F17]: Right now, the offset variable is the current size, not the max size from current surveys.

Median Imputation

If sufficient auxiliary data is not available, we will impute the pop size with median population within an imputation cell. This method involves partitioning the GQ universe into imputation cells based on the

detailed GQ type and GQ status. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median GQ pop size in the cell.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

We will evaluate the imputation methods using cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The remaining 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will impute the GQ pop size for all GQs in the validation deck. That is, we will attempt to impute the GQ pop size for every GQ in the 60% sample four times (once for each of the four methods). Then, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each method. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Some methods may perform better than others for certain types of units. For example, Poisson regression might perform best when the GQAC expected count is available, but not well when it is missing. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Appendix

Table 10: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Table 11: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Table 12: GQ Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Table 13: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Table 14: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

Background

Commented [JEZ(F1]: Tables based on 12/13/20 data.

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A telephone operation is in progress to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that open on Census Day, but vacant during the GQ Enumeration visit (which started in July 2020) require imputation.

In addition, we will impute a pop size for GQs that have a reported Census Day population count that is much smaller than expected. Our initial proposal is to impute when the Census Day population count is 25% of the GQAC expected count, but research into determining (and refining) this threshould is ongoing.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Visit but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with much lower than expected population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe.

Commented [PJC(F2]: For condition situation (2), do we have to mention other statuses?

Commented [JEZ(F3R2]: I think this question needs SME input. Are we accepting Vacant, Delete and Nonresidential GQ counts as-is?

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Open on Census Day, Vacant During Visit	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Commented [PJC(F4]: We should explain, perhaps insert a note below the table or in the text that differentiates "Vacant During Visit, Open on Census Day" from "Vacant GQ"? I presume the two are mutually exclusive. So did the former report that they were open on CD, while the latter didn't report anything about CD? I don't understand.

Commented [JEZ(F5R4]: Need to ask Debbie and Ryan.

Additionally, some of the 185,000 resolved occupied GQs will be treated as unresolved because their census day population is much lower than expected. The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs as well as any GQs with a much lower than expected population count. Our current threshold for a "low" population count is < 25% of the GQAC expected count. Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ status. Of the resolved GQs, 89,000 had a GQAC expected count and 90,000 did not. The

unresolved GQs include the 43,000 GQs without a reported count as well as 4,500 that had a large discrepancy between the GQAC expected population and the reported pop size.

Table 2: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Status	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	88,500	88,000	3,900	17,000	197,000
Vacant During Visit, Open on Census Day	1,000	550	300	19,500	21,500
Refusal GQ	350	450	300	6,700	7,800
Total	90,000	89,000	4,500	43,000	227,000

Commented [JEZ(F6]: Defined as pop count < 25% of expected. If the threshold is changed to < 10% of expected, count becomes 2,000.

Commented [JEZ(F7R6]: There also exist cases where the expected size is the same for all GQs in same facility. Sometimes these make sense, but sometimes it looks like they may be totals, when comparing to GP. For the unresolved, might not be able to tell.

Commented [JEZ(F8R6]: Might want to flag low count cases and do a manual review to determine if they may need imputation. Seems like expected count could have some measurement error issues, so we may not want to depend on it completely to determine if the CD pop is really too low.

Commented [JEZ(F9]: 100 of these have expected size <= 5. An additional 350 have expected size between 6 and 10.

Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 10 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Type	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	9,900	3,100	300	2,800	16,000
Juvenile Facilities	2,300	3,600	300	1,800	8,000
Nursing Facilities*	6,000	19,000	450	3,200	28,500
Hospitals	750	1,100	100	800	2,800
College Housing*	12,000	17,000	1,400	5,500	36,000
Military*	2,100	900	100	1,900	5,000
Shelters	21,000	3,200	550	8,200	33,000
Group Homes	29,000	32,500	850	9,100	72,000
Other	7,100	8,600	500	9,700	26,000
Total	90,000	89,000	4,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

An alternate or complimentary definition for a low census day population count would be to use 10% of the GQAC Max Number of People. Applying this definition would result in 1,100 more unresolved GQs, in addition to the 43,000 and 4,500 unresolved GQs in Table 3. Table 4 shows counts of the resolved and unresolved cases using this alternate threshold by GQ status. Table 5 shows the same information by GQ type.

Table 4: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Status	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	67,000	111,000	2,900	17,000	197,000
Vacant During Visit, Open on Census Day	550	1,000	350	19,500	21,500
Refusal GQ	150	650	300	6,700	7,800
Total	68,000	112,000	3,500	43,000	227,000

Note that 2,400 GQs with the Low Census Day Pop based on the Max Pop also have a Low Census Day Pop using the GQAC Expected Population.

Commented [JEZ(F10]: 2,400 GQs have < 25% of expected count and < 10% of max count.

Table 5: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Type	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	5,600	7,200	400	2,800	16,000
Juvenile Facilities	1,600	4,400	150	1,800	8,000
Nursing Facilities*	4,300	20,500	300	3,200	28,500
Hospitals	550	1,300	90	800	2,800
College Housing*	7,800	21,500	1,200	5,500	36,000
Military*	1,500	1500	90	1,900	5,000
Shelters	17,000	7,300	300	8,200	33,000
Group Homes	24,000	38,500	450	9,100	72,000
Other	5,600	10,000	450	9,700	26,000
Total	68,000	112,000	3,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Commented [PJC(F11]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Variables

Table 6 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, and Administrative Records. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 6: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR

Additional sources available for college housing GQs include data collected via web-scraping data from the Integrated Postsecondary Education Data System (IPEDS) and data from the Common Core. These variables are available at the facility level but not for individual MAFIDs.

Commented [JEZ(F12]: Is this only for 501s?

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) **"capacity utilization"**—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Additional facility-level variables may become available as research continues.

Table 7: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

First, if a pop count is available from the NPC call operation, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will use a combination of the following methods:

1. Ratio Imputation
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

Commented [ADK(F13]: Need to look at paradata as covariates as well on the models

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 8 shows that 8,610 of the unresolved GQ can be resolved by converting the GQAC expected count to the GQ pop count using the following ratio adjustment.

Table 8: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For each GQ type, we will use the ratio of the reported GQ Census Day count to the GQAC expected count to convert the GQAC expected count of the unresolved GQ to a Census Day imputed count. For each GQ type, we will calculate the ratio of the sum of the GQAC Expected Count to the sum of the reported GQ population for the resolved cases. For the unresolved GQs, we will multiply the GQAC expected count by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = \ GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. We will not use ratio imputation with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count. Conversion factors for the four variables under consideration are

shown in Table 9. Table 12Table 14 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 9: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0 3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1.017	0.6385
College Housing	0.7818	0.5492	0.9444	0 5535
Military	0.7317	0.2290	0.9492	0 2914
Shelters	0.6261	0.5325	0.6180	0 5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0 3597
All GQs	0.7878	0.5057	0.9217	0 5153

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for GQs for colleges and universities (GQTYPCUR=501).

Adjusting the IPEDS facility-level Room Capacity

To adjust the IPEDS room capacity for reference year differences, we use the GQAC Max Number of People. We first select colleges for which we have a positive GQAC Max Number of People for every GQ at the facility. Since the IPEDS data does not include off-campus housing, we further subset on facilities that have no Greek letter GQs (fraternity or sorority houses). Finally, to maximize the chances that we are comparing apples to apples, we also subset to facilities for which the match quality is very high (match score > 90%). Within this subset, we calculate the average ratio of the facility-level sum of GQAC Max Number of People over the room capacity from IPEDS:

$$Average\ Ratio_S \ = \sum_{i \in S} \frac{\sum_{college_i} GQAC\ Max\ Number\ of\ People}{IPEDS\ Room\ Capacity\ at\ college\ i}$$

where *S* is the set of colleges with no Greek GQs only positive values for GQAC Max Number of People.

Reassuringly, within this set of colleges, the median ratio is [redacted], the mode is [redacted], the 25th percentile is [redacted], and the 75th percentile is [redacted].

After adjusting the IPEDS college-level room capacity, we will similarly adjust for GQ "capacity utilization" at the college-level, using the mean ratio of 2020 Census Day GQ population over GQAC Max Number of People for all GQs for which both 2020 Census Day GQ population over GQAC Max Number of People. If time and sample sizes permit, we will also calculate this average ratio for college size classes. If the mean ratios differ significantly by college size class we will use separate capacity utilization adjustment for each college size class

Commented [TLK(F14]: I think a word is missing from this sentence.

After adjusting the college-level total room capacity to account reference year for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Finally, we will adjust the room capacity for GQ population in off-campus Greek housing (which is not included in the IPEDS room capacity). About 51% of colleges in the GQ data have no Greek letter GQs. However, among colleges with at least 1 Greek letter GQ, at the mean has 38% of GQs are Greek letter houses, with a standard deviation of 34%. Since the importance of Greek letter GQs varies widely across colleges, we apply a Greek housing adjustment to each college based on which of 5 categories the colleges falls into:

Commented [TLK(F15]: Clarify this.

1. No Greek housing GQs
2. Small school, low percentage of Greek housing GQs
3. Small school, high percentage of Greek housing GQs
4. Large school, low percentage of Greek housing GQs
5. Small school, high percentage of Greek housing GQs

For colleges with no/low GQ missingness rates, we take the average within each category of Greek housing pop counts over total GQ pop counts.

Commented [TLK(F16]: Uncluer what the average is of.

Once we calculate the collete-level residual, we will then allocate the population counts among the GQs in the college.

Modeling

A third approach would be to impute the GQ pop counts from a Poisson regression model. The dependent variable will be reported GQ pop count with an offset of the max number of people (because that is filled the most). Independent variables will be selected from Table 6. It is important to note that GQ type will either be a fixed-effect covariate in the models or separate models will be fit by GQ type. Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

Commented [ADK(F17]: Right now, the offset variable is the current size, not the max size from current surveys.

Median Imputation

If sufficient auxiliary data is not available, we will impute the pop size with median population within an imputation cell. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and GQ status. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median GQ pop size in the cell.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

We will evaluate the imputation methods using cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The remaining 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will impute the GQ pop size for all GQs in the validation deck. That is, we will attempt to impute the GQ pop size for every GQ in the 60% sample four times (once for each of the four methods). Then, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each method. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Some methods may perform better than others for certain types of units. For example, Poisson regression might perform best when the GQAC expected count is available, but not well when it is missing. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Appendix

Table 10: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Table 11: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Table 12: GQ Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Table 13: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Table 14: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A telephone operation is in progress to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that open on Census Day, but vacant during the GQ Enumeration visit (which started in July 2020) require imputation.

In addition, we will impute a pop size for GQs that have a reported Census Day population count that is much smaller than expected. Our initial proposal is to impute when the Census Day population count is 25% of the GQAC expected count, but research into determining (and refining) this threshould is ongoing.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Visit but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with much lower than expected population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe.

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Open on Census Day, Vacant During Visit	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Additionally, some of the 185,000 resolved occupied GQs will be treated as unresolved because their census day population is much lower than expected. The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs as well as any GQs with a much lower than expected population count. Our current threshold for a “low” population count is < 25% of the GQAC expected count. Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ status. Of the resolved GQs, 89,000 had a GQAC expected count and 90,000 did not. The

unresolved GQs include the 43,000 GQs without a reported count as well as 4,500 that had a large discrepancy between the GQAC expected population and the reported pop size.

Table 2: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Status	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	88,500	88,000	3,900	17,000	197,000
Vacant During Visit, Open on Census Day	1,000	550	300	19,500	21,500
Refusal GQ	350	450	300	6,700	7,800
Total	90,000	89,000	4,500	43,000	227,000

Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 10 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Type	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	9,900	3,100	300	2,800	16,000
Juvenile Facilities	2,300	3,600	300	1,800	8,000
Nursing Facilities*	6,000	19,000	450	3,200	28,500
Hospitals	750	1,100	100	800	2,800
College Housing*	12,000	17,000	1,400	5,500	36,000
Military*	2,100	900	100	1,900	5,000
Shelters	21,000	3,200	550	8,200	33,000
Group Homes	29,000	32,500	850	9,100	72,000
Other	7,100	8,600	500	9,700	26,000
Total	90,000	89,000	4,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

An alternate definition for a low census day population count would be to use 10% of the GQAC Max Number of People. **Error! Not a valid bookmark self-reference.** shows counts of the resolved and unresolved cases using this alternate threshold by GQ status. Table 5 shows the same information by GQ type. We will examine using the intersection or union of these conditions as well as setting thresholds at different levels to determine which reported counts require imputation.

Table 4: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Status	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	67,000	111,000	2,900	17,000	197,000
Vacant During Visit, Open on Census Day	550	1,000	350	19,500	21,500
Refusal GQ	150	650	300	6,700	7,800
Total	68,000	112,000	3,500	43,000	227,000

Note that 2,400 GQs with the Low Census Day Pop based on the Max Pop also have a Low Census Day Pop using the GQAC Expected Population.

Table 5: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Type	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	5,600	7,200	400	2,800	16,000
Juvenile Facilities	1,600	4,400	150	1,800	8,000
Nursing Facilities*	4,300	20,500	300	3,200	28,500
Hospitals	550	1,300	90	800	2,800
College Housing*	7,800	21,500	1,200	5,500	36,000
Military*	1,500	1500	90	1,900	5,000
Shelters	17,000	7,300	300	8,200	33,000
Group Homes	24,000	38,500	450	9,100	72,000
Other	5,600	10,000	450	9,700	26,000
Total	68,000	112,000	3,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Variables

Table 6 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, and Administrative Records. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 6: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR

Additional sources available for college housing GQs include data collected via web-scraping, data from the Integrated Postsecondary Education Data System (IPEDS) and data from the Common Core. These variables are available at the facility level but not for individual MAFIDs.

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) “**capacity utilization**”—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Additional facility-level variables may become available as research continues.

Table 7: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

First, if a pop count is available from the NPC call operation, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will use a combination of the following methods:

1. Ratio Imputation
2. Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 8 shows that 8,600 of the unresolved GQ can be resolved by converting the GQAC expected count to the GQ pop count using the following ratio adjustment.

Table 8: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For each GQ type, we will use the ratio of the reported GQ Census Day count to the GQAC expected count to convert the GQAC expected count of the unresolved GQ to a Census Day imputed count. For each GQ type, we will calculate the ratio of the sum of the GQAC Expected Count to the sum of the reported GQ population for the resolved cases. For the unresolved GQs, we will multiply the GQAC expected count by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. We will not use ratio imputation with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count. Conversion factors for the four variables under consideration are

shown in Table 9. Table 12Table 14 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 9: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0.3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1.017	0.6385
College Housing	0.7818	0.5492	0.9444	0.5535
Military	0.7317	0.2290	0.9492	0.2914
Shelters	0.6261	0.5325	0.6180	0.5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0.3597
All GQs	0.7878	0.5057	0.9217	0.5153

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for GQs for colleges and universities (GQTYPCUR=501).

First, we will adjust the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population, GQAC Max Number of People, and Greek Housing variables.

After adjusting the college-level total room capacity to account reference year and for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Once we calculate the college-level residual, we will then allocate the population counts among the GQs in the college without GQAC Expected Count.

Modeling

A third approach would be to impute the GQ pop counts from a Poisson regression model. The dependent variable will be reported GQ pop count with an offset of the max number of people (because that is filled the most). Independent variables will be selected from Table 6. It is important to note that GQ type will either be a fixed-effect covariate in the models or separate models will be fit by GQ type.

Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

Median Imputation

If sufficient auxiliary data is not available, we will impute the pop size with median population within an imputation cell. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and GQ status. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median GQ pop size in the cell.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

We will evaluate the imputation methods using cross validation. First, we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The remaining 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will impute the GQ pop size for all GQs in the validation deck. That is, we will attempt to impute the GQ pop size for every GQ in the 60% sample four times (once for each of the four methods). Note that the second method can only be applied to college housing. Then, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each method. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Some methods may perform better than others for certain types of units. For example, Poisson regression might perform best when the GQAC expected count is available, but not well when it is missing. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Appendix

Table 10: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Table 11: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Table 12: GQ Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Table 13: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Table 14: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A telephone operation is in progress to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that open on Census Day, but vacant during the GQ Enumeration visit (which started in July 2020) require imputation.

In addition, we will impute a pop size for GQs that have a reported Census Day population count that is much smaller than expected. Our initial proposal is to impute when the Census Day population count is 25% of the GQAC expected count, but research into determining (and refining) this threshould is ongoing.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Visit but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with much lower than expected population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe.

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Open on Census Day, Vacant During Visit	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Additionally, some of the 185,000 resolved occupied GQs will be treated as unresolved because their census day population is much lower than expected. The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs as well as any GQs with a much lower than expected population count. Our current threshold for a "low" population count is < 25% of the GQAC expected count. Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ status. Of the resolved GQs, 89,000 had a GQAC expected count and 90,000 did not. The

Commented [JEZ(F1]: Tables based on 12/13/20 data.

Commented [PJC(F2]: For condition situation (2), do we have to mention other statuses?

Commented [JEZ(F3R2]: I think this question needs SME input. Are we accepting Vacant, Delete and Nonresidential GQ counts as-is?

Commented [PJC(F4]: We should explain, perhaps insert a note below the table or in the text that differentiates "Vacant During Visit, Open on Census Day" from "Vacant GQ"? I presume the two are mutually exclusive. So did the former report that they were open on CD, while the latter didn't report anything about CD? I don't understand.

Commented [JEZ(F5R4]: Need to ask Debbie and Ryan.

unresolved GQs include the 43,000 GQs without a reported count as well as 4,500 that had a large discrepancy between the GQAC expected population and the reported pop size.

Table 2: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Status	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	88,500	88,000	3,900	17,000	197,000
Vacant During Visit, Open on Census Day	1,000	550	300	19,500	21,500
Refusal GQ	350	450	300	6,700	7,800
Total	90,000	89,000	4,500	43,000	227,000

Commented [JEZ(F6]: Defined as pop count < 25% of expected. If the threshold is changed to < 10% of expected, count becomes 2,000.

Commented [JEZ(F7R6]: There also exist cases where the expected size is the same for all GQs in same facility. Sometimes these make sense, but sometimes it looks like they may be totals, when comparing to GP. For the unresolved, might not be able to tell.

Commented [JEZ(F8R6]: Might want to flag low count cases and do a manual review to determine if they may need imputation. Seems like expected count could have some measurement error issues, so we may not want to depend on it completely to determine if the CD pop is really too low.

Commented [JEZ(F9]: 100 of these have expected size <= 5. An additional 350 have expected size between 6 and 10.

Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 10 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Type	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	9,900	3,100	300	2,800	16,000
Juvenile Facilities	2,300	3,600	300	1,800	8,000
Nursing Facilities*	6,000	19,000	450	3,200	28,500
Hospitals	750	1,100	100	800	2,800
College Housing*	12,000	17,000	1,400	5,500	36,000
Military*	2,100	900	100	1,900	5,000
Shelters	21,000	3,200	550	8,200	33,000
Group Homes	29,000	32,500	850	9,100	72,000
Other	7,100	8,600	500	9,700	26,000
Total	90,000	89,000	4,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

An alternate or complimentary definition for a low census day population count would be to use 10% of the GQAC Max Number of People. Applying this definition would result in 1,100 more unresolved GQs, in addition to the 43,000 and 4,500 unresolved GQs in Table 3. Table 4 shows counts of the resolved and unresolved cases using this alternate threshold by GQ status. Table 5 shows the same information by GQ type.

Commented [PJC(F10]: Sp. of "complementary." Also, by complementary, do you mean the intersection (both conditions must hold) or union (either)?

Commented [PJC(F11]: "... result in 1,100 more unresolved GQs ..." This implies you're using the Max Number as complementary (union) to the Expected Number. But in Tables 4 and 5, you have fewer unresolved, implying you're only using Max Number, or the intersection, not both. One could use Max Number as an alternate or as complementary. Or you could show tables for both. But currently it appears to be inconsistent.

Table 4: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Status	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	67,000	111,000	2,900	17,000	197,000
Vacant During Visit, Open on Census Day	550	1,000	350	19,500	21,500
Refusal GQ	150	650	300	6,700	7,800
Total	68,000	112,000	3,500	43,000	227,000

Note that 2,400 GQs with the Low Census Day Pop based on the Max Pop also have a Low Census Day Pop using the GQAC Expected Population.

Commented [JEZ(F12]: 2,400 GQs have < 25% of expected count and < 10% of max count.

Table 5: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Type	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	5,600	7,200	400	2,800	16,000
Juvenile Facilities	1,600	4,400	150	1,800	8,000
Nursing Facilities*	4,300	20,500	300	3,200	28,500
Hospitals	550	1,300	90	800	2,800
College Housing*	7,800	21,500	1,200	5,500	36,000
Military*	1,500	1500	90	1,900	5,000
Shelters	17,000	7,300	300	8,200	33,000
Group Homes	24,000	38,500	450	9,100	72,000
Other	5,600	10,000	450	9,700	26,000
Total	68,000	112,000	3,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Commented [PJC(F13]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Variables

Table 6 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, and Administrative Records. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 6: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR

Additional sources available for college housing GQs include data collected via web-scraping, data from the Integrated Postsecondary Education Data System (IPEDS) and data from the Common Core. These variables are available at the facility level but not for individual MAFIDs.

Commented [JEZ(F14]: Is this only for 501s?

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) **"capacity utilization"**—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**—IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Additional facility-level variables may become available as research continues.

Table 7: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

First, if a pop count is available from the NPC call operation, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will use a combination of the following methods:

1. Ratio Imputation
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

Commented [ADK(F15]: Need to look at paradata as covariates as well on the models

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 8 shows that 8,600 of the unresolved GQ can be resolved by converting the GQAC expected count to the GQ pop count using the following ratio adjustment.

Table 8: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For each GQ type, we will use the ratio of the reported GQ Census Day count to the GQAC expected count to convert the GQAC expected count of the unresolved GQ to a Census Day imputed count. For each GQ type, we will calculate the ratio of the sum of the GQAC Expected Count to the sum of the reported GQ population for the resolved cases. For the unresolved GQs, we will multiply the GQAC expected count by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. We will not use ratio imputation with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count. Conversion factors for the four variables under consideration are

shown in Table 9. Tables 12-14 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 9: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0.3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1.017	0.6385
College Housing	0.7818	0.5492	0.9444	0.5535
Military	0.7317	0.2290	0.9492	0.2914
Shelters	0.6261	0.5325	0.6180	0.5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0.3597
All GQs	0.7878	0.5057	0.9217	0.5153

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for GQs for colleges and universities (GQTYPCUR=501).

~~Adjusting the IPEDS facility-level Room Capacity~~

~~To adjust the IPEDS room capacity for reference year differences, we use the GQAC Max Number of People. We first select colleges for which we have a positive GQAC Max Number of People for every GQ at the facility. Since the IPEDS data does not include off-campus housing, we further subset on facilities that have no Greek letter GQs (fraternity or sorority houses). Finally, to maximize the chances that we are comparing apples to apples, we also subset to facilities for which the match quality is very high (match score > 90%). Within this subset, we calculate the average ratio of the facility-level sum of GQAC Max Number of People over the room capacity from IPEDS:~~

~~$Average\ Ratio_S = \sum_{i \in S} \frac{\sum_{college_i} GQAC\ Max\ Number\ of\ People}{IPEDS\ Room\ Capacity\ at\ college\ i}$~~

~~where S is the set of colleges with no Greek GQs only positive values for GQAC Max Number of People.~~

~~Reassuringly, within this set of colleges, the median ratio is~~ [redacted] ~~, the mode is~~ [redacted] ~~the 25th percentile is~~ [redacted] ~~, and the 75th percentile is~~ [redacted] ~~.~~

~~After adjusting the IPEDS college-level room capacity, we will similarly adjust for GQ "capacity utilization" at the college-level, using the mean ratio of 2020 Census Day GQ population over GQAC Max Number of People for all GQs for which both 2020 Census Day GQ population over GQAC Max Number of People. If time and sample sizes permit, we will also calculate this average ratio for college size classes. If the mean ratios differ significantly by college size class we will use separate capacity utilization adjustment for each college size class~~

Commented [TLK(F16]: I think a word is missing from this sentence.

Commented [JEZ(F17]: Good info but removing for now to keep this more high-level.

First, we will adjust the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population, GQAC Max Number of People, and Greek Housing variables.

Commented [JEZ(F18]: Does this summary make sense?

Commented [TKW(F19]: Yes, this makes sense.

After adjusting the college-level total room capacity to account reference year and for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

~~Finally, we will adjust the room capacity for GQ population in off-campus Greek housing (which is not included in the IPEDS room capacity). About 51% of colleges in the GQ data have no Greek letter GQs. However, among colleges with at least 1 Greek letter GQ, at the mean has 38% of GQs are Greek letter houses, with a standard deviation of 34%. Since the importance of Greek letter GQs varies widely across colleges, we apply a Greek housing adjustment to each college based on which of 5 categories the colleges falls into:~~

Commented [TLK(F20]: Clarify this.

1. ~~No Greek housing GQs~~
2. ~~Small school, low percentage of Greek housing GQs~~
3. ~~Small school, high percentage of Greek housing GQs~~
4. ~~Large school, low percentage of Greek housing GQs~~
5. ~~Small school, high percentage of Greek housing GQs~~

~~For colleges with no/low GQ missingness rates, we take the average within each category of Greek housing pop counts over total GQ pop counts.~~

Commented [TLK(F21]: Unclear what the average is of.

Once we calculate the college-level residual, we will then allocate the population counts among the GQs in the college without GQAC Expected Count.

Commented [JEZ(F22]: Is this right?

Commented [TKW(F23]: Yes, this is right.

Modeling

A third approach would be to impute the GQ pop counts from a Poisson regression model. The dependent variable will be reported GQ pop count with an offset of the max number of people (because that is filled the most). Independent variables will be selected from Table 6. It is important to note that GQ type will either be a fixed-effect covariate in the models or separate models will be fit by GQ type. Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

Commented [ADK(F24]: Right now, the offset variable is the current size, not the max size from current surveys.

Median Imputation

If sufficient auxiliary data is not available, we will impute the pop size with median population within an imputation cell. This method involves partitioning the GQ universe into imputation cells based on the

detailed GQ type and GQ status. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median GQ pop size in the cell.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

We will evaluate the imputation methods using cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The remaining 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will impute the GQ pop size for all GQs in the validation deck. That is, we will attempt to impute the GQ pop size for every GQ in the 60% sample four times (once for each of the four methods). Then, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each method. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Some methods may perform better than others for certain types of units. For example, Poisson regression might perform best when the GQAC expected count is available, but not well when it is missing. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Appendix

Table 10: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Table 11: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Table 12: GQ Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Table 13: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Table 14: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Background

Commented [JEZ(F1]: Tables based on 12/13/20 data.

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A telephone operation is in progress to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that open on Census Day, but vacant during the GQ Enumeration visit (which started in July 2020) require imputation.

In addition, we will impute a pop size for GQs that have a reported Census Day population count that is much smaller than expected. Our initial proposal is to impute when the Census Day population count is 25% of the GQAC expected count, but research into determining (and refining) this threshould is ongoing.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Visit but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with much lower than expected population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe.

Commented [PJC(F2]: For condition situation (2), do we have to mention other statuses?

Commented [JEZ(F3R2]: I think this question needs SME input. Are we accepting Vacant, Delete and Nonresidential GQ counts as-is?

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Open on Census Day, Vacant During Visit	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Commented [JEZ(F4]: From Diedre: Could you please confirm that you are thinking expected to be occupied based on information collected during GQAC? As the only source of your decision?

I think we need to explain where we get this status, related to Pat's question below.

Commented [PJC(F5]: We should explain, perhaps insert a note below the table or in the text that differentiates "Vacant During Visit, Open on Census Day" from "Vacant GQ"? I presume the two are mutually exclusive. So did the former report that they were open on CD, while the latter didn't report anything about CD? I don't understand.

Commented [JEZ(F6R5]: Need to ask Debbie and Ryan.

Additionally, some of the 185,000 resolved occupied GQs will be treated as unresolved because their census day population is much lower than expected. The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs as well as any GQs with a much lower than expected population count. Our current threshold for a "low" population count is < 25% of the GQAC expected count. Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ status. Of the resolved GQs, 89,000 had a GQAC expected count and 90,000 did not. The

unresolved GQs include the 43,000 GQs without a reported count as well as 4,500 that had a large discrepancy between the GQAC expected population and the reported pop size.

Table 2: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

	Resolved		Unresolved		
GQ Status	**No Expected Pop**	**Reasonable Census Day Pop**	**Low Census Day Pop**	**No Census Day Pop**	**Total**
Occupied GQ	88,500	88,000	3,900	17,000	197,000
Vacant During Visit, Open on Census Day	1,000	550	300	19,500	21,500
Refusal GQ	350	450	300	6,700	7,800
Total	90,000	89,000	4,500	43,000	227,000

Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 10 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

	Resolved		Unresolved		
GQ Type	**No Expected Pop**	**Reasonable Census Day Pop**	**Low Census Day Pop**	**No Census Day Pop**	**Total**
Correctional Facilities*	9,900	3,100	300	2,800	16,000
Juvenile Facilities	2,300	3,600	300	1,800	8,000
Nursing Facilities*	6,000	19,000	450	3,200	28,500
Hospitals	750	1,100	100	800	2,800
College Housing*	12,000	17,000	1,400	5,500	36,000
Military*	2,100	900	100	1,900	5,000
Shelters	21,000	3,200	550	8,200	33,000
Group Homes	29,000	32,500	850	9,100	72,000
Other	7,100	8,600	500	9,700	26,000
Total	90,000	89,000	4,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

An alternate ~~or complimentary~~ definition for a low census day population count would be to use 10% of the GQAC Max Number of People. ~~Applying this definition would result in 1,100 more unresolved GQs, in addition to the 43,000 and 4,500 unresolved GQs in Table 3.~~ Table 4 shows counts of the resolved and unresolved cases using this alternate threshold by GQ status. Table 5 shows the same information by GQ type. We will examine using the intersection or union of these conditions as well as setting thresholds at different levels to determine which reported counts require imputation.

Commented [JEZ(F7]: Defined as pop count < 25% of expected. If the threshold is changed to < 10% of expected, count becomes 2,000.

Commented [JEZ(F8R7]: There also exist cases where the expected size is the same for all GQs in same facility. Sometimes these make sense, but sometimes it looks like they may be totals, when comparing to GP. For the unresolved, might not be able to tell.

Commented [JEZ(F9R7]: Might want to flag low count cases and do a manual review to determine if they may need imputation. Seems like expected count could have some measurement error issues, so we may not want to depend on it completely to determine if the CD pop is really too low.

Commented [JEZ(F10]: 100 of these have expected size <= 5. An additional 350 have expected size between 6 and 10.

Commented [JEZ(F11]: [12/15/2020 4:30 PM] Deborah Stempowski (CENSUS/ADDC FED): Looks like All GQ Types 200s (Juvenile items) was about 9600 GQs sent to enumeration

[12/15/2020 4:32 PM] Michael R Ratcliffe (CENSUS/GEO FED):
We had 9,700 juvenile facilities GQs in 2010 (4% of all GQs).

Commented [JEZ(F12]: From James Christy: tricky to split out facility level counts for GQs. We might consider to looking at facility-level data to determine where we need to do imputation.

Table 4: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Status	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	67,000	111,000	2,900	17,000	197,000
Vacant During Visit, Open on Census Day	550	1,000	350	19,500	21,500
Refusal GQ	150	650	300	6,700	7,800
Total	68,000	112,000	3,500	43,000	227,000

Note that 2,400 GQs with the Low Census Day Pop based on the Max Pop also have a Low Census Day Pop using the GQAC Expected Population.

Table 5: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Type	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	5,600	7,200	400	2,800	16,000
Juvenile Facilities	1,600	4,400	150	1,800	8,000
Nursing Facilities*	4,300	20,500	300	3,200	28,500
Hospitals	550	1,300	90	800	2,800
College Housing*	7,800	21,500	1,200	5,500	36,000
Military*	1,500	1500	90	1,900	5,000
Shelters	17,000	7,300	300	8,200	33,000
Group Homes	24,000	38,500	450	9,100	72,000
Other	5,600	10,000	450	9,700	26,000
Total	68,000	112,000	3,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Variables

Table 6 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, and Administrative Records. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Commented [JEZ(F13]: 2,400 GQs have < 25% of expected count and < 10% of max count.

Commented [JEZ(F14]: Deb's question about Shelters – includes SBEs? I didn't hear the acronym. Pat said yes.

Commented [JEZ(F15R14]: From Al: How will we know when we should be imputing for SBEs? SBEs may be closed.

From Al: Are we only imputing for the asterisks types?

Pat: We are starting with those, if we think our models work we will decide if we want to impute for other types.

Commented [JEZ(F16R14]: Tori: What about juvenile facilities? How much of the 2010 GQ pop? Should we even impute for them?
Karen: we can look into this.

In general, think about contribution of different GQ types to overall GQ universe before deciding what we want to impute vs accepting the zero.

Commented [PJC(F17]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Table 6: Auxiliary and Historical Data at the GQ-Level

Commented [JEZ(F18]: From chat in EGG meeting: use 5-year ACS estimates?

Commented [JEZ(F19R18]: From James Christy: FWIW - when we visit a GQ for ACS, we ask for the total population of that GQ, then use that to sub-sample for selecting cases for interview. Reference date is when we visit the GQ. (Unlike Decennial which has a fixed reference date). I don't think that total pop count is part of what's published - but could be wrong.

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR

Additional sources available for college housing GQs include data collected via web-scraping data from the Integrated Postsecondary Education Data System (IPEDS) and data from the Common Core. These variables are available at the facility level but not for individual MAFIDs.

Commented [JEZ(F20]: Is this only for 501s?

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) **"capacity utilization"**—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Additional facility-level variables may become available as research continues.

Table 7: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

First, if a pop count is available from the NPC call operation, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will use a combination of the following methods:

1. Ratio Imputation
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

Commented [ADK(F21]: Need to look at paradata as covariates as well on the models

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 8 shows that 8,600 of the unresolved GQ can be resolved by converting the GQAC expected count to the GQ pop count using the following ratio adjustment.

Table 8: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For each GQ type, we will use the ratio of the reported GQ Census Day count to the GQAC expected count to convert the GQAC expected count of the unresolved GQ to a Census Day imputed count. For each GQ type, we will calculate the ratio of the sum of the GQAC Expected Count to the sum of the reported GQ population for the resolved cases. For the unresolved GQs, we will multiply the GQAC expected count by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. We will not use ratio imputation with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count. Conversion factors for the four variables under consideration are

shown in Table 9. Table 12Table 14 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 9: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0 3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1.017	0.6385
College Housing	0.7818	0.5492	0.9444	0 5535
Military	0.7317	0.2290	0.9492	0 2914
Shelters	0.6261	0.5325	0.6180	0 5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0 3597
All GQs	0.7878	0.5057	0.9217	0 5153

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for GQs for colleges and universities (GQTYPCUR=501).

~~***Adjusting the IPEDS facility-level Room Capacity***~~

~~To adjust the IPEDS room capacity for reference year differences, we use the GQAC Max Number of People. We first select colleges for which we have a positive GQAC Max Number of People for every GQ at the facility. Since the IPEDS data does not include off-campus housing, we further subset on facilities that have no Greek letter GQs (fraternity or sorority houses). Finally, to maximize the chances that we are comparing apples to apples, we also subset to facilities for which the match quality is very high (match score > 90%). Within this subset, we calculate the average ratio of the facility-level sum of GQAC Max Number of People over the room capacity from IPEDS:~~

~~$Average\ Ratio_S = \sum_{i \in S} \frac{\sum_{H \in g_i} GQAC\ Max\ Number\ of\ People}{IPEDS\ Room\ Capacity\ at\ college\ i}$~~

~~where S is the set of colleges with no Greek GQs only positive values for GQAC Max Number of People.~~

~~Reassuringly, within this set of colleges, the median ratio is~~ [redacted]~~, the mode is~~ [redacted]~~, the 25th percentile is~~ [redacted]~~, and the 75th percentile is~~ [redacted]~~.~~

~~After adjusting the IPEDS college-level room capacity, we will similarly adjust for GQ "capacity utilization" at the college level, using the mean ratio of 2020 Census Day GQ population over GQAC Max Number of People for all GQs for which both 2020 Census Day GQ population over GQAC Max Number of People. If time and sample sizes permit, we will also calculate this average ratio for college size classes. If the mean ratios differ significantly by college size class we will use separate capacity utilization adjustment for each college size class~~

Commented [TLK(F22]: I think a word is missing from this sentence.

Commented [JEZ(F23]: Good info but removing for now to keep this more high-level.

First, we will adjust the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population, GQAC Max Number of People, and Greek Housing variables.

Commented [JEZ(F24]: Does this summary make sense?

Commented [TKW(F25]: Yes, this makes sense.

After adjusting the college-level total room capacity to account reference year and for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

~~Finally, we will adjust the room capacity for GQ population in off-campus Greek housing (which is not included in the IPEDS room capacity). About 51% of colleges in the GQ data have no Greek letter GQs. However, among colleges with at least 1 Greek letter GQ, at the mean has 38% of GQs are Greek letter houses, with a standard deviation of 34%. Since the importance of Greek letter GQs varies widely across colleges, we apply a Greek housing adjustment to each college based on which of 5 categories the colleges falls into:~~

Commented [TLK(F26]: Clarify this.

1. ~~No Greek housing GQs~~
2. ~~Small school, low percentage of Greek housing GQs~~
3. ~~Small school, high percentage of Greek housing GQs~~
4. ~~Large school, low percentage of Greek housing GQs~~
5. ~~Small school, high percentage of Greek housing GQs~~

~~For colleges with no/low GQ missingness rates, we take the average within each category of Greek housing pop counts over total GQ pop counts.~~

Commented [TLK(F27]: Uncluer what the average is of.

Once we calculate the college-level residual, we will then allocate the population counts among the GQs in the college without GQAC Expected Count.

Commented [JEZ(F28]: Is this right?

Commented [TKW(F29]: Yes, this is right.

Modeling

A third approach would be to impute the GQ pop counts from a Poisson regression model. The dependent variable will be reported GQ pop count with an offset of the max number of people (because that is filled the most). Independent variables will be selected from Table 6. It is important to note that GQ type will either be a fixed-effect covariate in the models or separate models will be fit by GQ type. Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

Commented [ADK(F30]: Right now, the offset variable is the current size, not the max size from current surveys.

Median Imputation

If sufficient auxiliary data is not available, we will impute the pop size with median population within an imputation cell. This method involves partitioning the GQ universe into imputation cells based on the

detailed GQ type and GQ status. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median GQ pop size in the cell.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

We will evaluate the imputation methods using cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The remaining 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will impute the GQ pop size for all GQs in the validation deck. That is, we will attempt to impute the GQ pop size for every GQ in the 60% sample four times (once for each of the four methods). Then, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each method. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Some methods may perform better than others for certain types of units. For example, Poisson regression might perform best when the GQAC expected count is available, but not well when it is missing. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Commented [JEZ(F31]: From John Abowd: why was 40/60 training/validation single train/validation selected over 50/50 cross-validation?

Commented [JEZ(F32R31]: Leave-out one estimators.

Appendix

Table 10: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Table 11: GQAC Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Table 12: GQAC Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Table 13: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Table 14: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

This Document Contains Title-13 Data

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

Background

Commented [JEZ(F1]: Tables based on 12/18/20 data.

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A special telephone operation was conducted to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation that pass an initial quality review as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that open on Census Day, but vacant during the GQ Enumeration visit (which started in July 2020) require imputation.

In addition, we will impute a pop size for GQs that do not meet our quality edits. These GQs have a reported population that is not plausible and likely a error. For example, a facility might have reported their facilty population in one GQ. We employed the Hidiroglou-Berthelot (HB) editing process, commonly used in establishment surveys, to identify these GQs with an implausible reported population. Information about HB edits is included in the Appendix.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have an unplausible reported count. This universe is made up of GQs with a status of Occupied; Open on Census Day, but Vacant During Visit[1]; and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with implausible population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe. The other statuses of Vacant, Delete, and Nonresidential are considered out-of-scope for GQ Count Imputation. The GQ Count Imputaiton will only impute a positive population.

Table 1: GQ Universe

Commented [JEZ(F2]: Do we want to do any tables about the GEO review? Or will we just re-create this table after we incorporate the 'N's as resolved?

Commented [TLK(F3R2]: I think this table should be updated once the review is complete. Thus the N will be included in the resolved column.

Commented [PJC(F4]: I agree. And somewhere we can describe briefly the calling operation, the data collected in it, and GEO's review.

GQ Status	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Occupied GQ	177,000	17,000	3,100	197,000
Open on Census Day, Vacant During Visit	1,700	19,500	200	21,500
Refusal GQ	900	6,700	200	7,800
Vacant GQ	30,500	0	0	30,500
Delete GQ	7,600	0	0	7,600
Nonresidential GQ	2,500	0	0	2,500
Total	220,000	43,000	3,500	267,000

[1] During GQ enumeration, the GQ was found to be vacant, but the contact at the GQ said the GQ was open on Census Day. This is different from the vacant GQs which were reported to be vacant on Census Day.

Some of the occupied GQs with a reported population were treated as unresolved because their census day population was implausible. The goal of the GQ Count Imputation is to determine a population count for all 43,000 occupied GQs with no reported population as well as the 3,500 occupied GQs with an implausible population count.

Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 7 in the Appendix has a full list of the GQ type codes.

Table 2: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

Commented [PJC(F5]: We should insert "occupied GQs" or something similar (occupied or suspected occupied) in the table title somewhere.

GQ Type	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Correctional Facilities*	13,000	2,800	250	16,000
Juvenile Facilities	6,100	1,800	90	8,000
Nursing Facilities*	24,500	3,200	550	28,500
Hospitals	1,900	800	70	2,800
College Housing*	29,000	5,500	1,300	36,000
Military*	3,000	1,900	70	5,000
Shelters	24,500	8,200	150	33,000
Group Homes	62,000	9,100	700	72,000
Other	16,000	9,700	300	26,000
Total	180,000	43,000	3,500	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Commented [PJC(F6]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Added: As our method has developed, and we're embracing the median imputation option when no data are available, it appears that this set will be empty.

Variables

Table 3 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, Administrative Records, and nursing home data from the Centers for Medicare & Medicaid Services (CMS). We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 3: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR
Number of All Beds	Total number of resident beds in the facility as reported by the provider.	CMS
Number of Occupied Beds	Total number of resident beds that are currently occupied as reported by the provider.	CMS

Commented [JEZ(F7]: From chat in EGG meeting: use 5-year ACS estimates?

Commented [JEZ(F8R7]: From James Christy: FWIW - when we visit a GQ for ACS, we ask for the total population of that GQ, then use that to sub-sample for selecting cases for interview. Reference date is when we visit the GQ. (Unlike Decennial which has a fixed reference date). I don't think that total pop count is part of what's published - but could be wrong.

Commented [TLK(F9R7]: Stuart Irby confirmed what James says. The Current GQ Size contains the size when conducting the listing. Current Surveys also updates the MAF in the same way as ACS.

Commented [JEZ(F10]: Add nursing home data from CMS. MEPS data?

Additional sources available for college housing GQs include data collected via web-scraping and data from the Integrated Postsecondary Education Data System (IPEDS). These variables are available at the facility level but not for individual MAFIDs.

Commented [JEZ(F11]: Is this only for 501s?

Commented [JEZ(F12R11]: I'm not sure now how we will use these data. It seems like they are most useful for determining vacant or delete status. I don't know ifw e c

Commented [TLK(F13R11]: That's right. Unless they can get pop counts soon, we won't be using the web-scraping data.

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) "**capacity utilization**"—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Additional facility-level variables may become available as research continues.

Table 4: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Data Editing

The HB edits employed to detect implausible pop counts will also be used to determine which resolved GQs contribute to the donor pool for imputation. While the most extreme outliers are flagged for imputation, for less extreme outliers, we will accept the reported values, but keep those GQs from contributing to the imputation. We compared reported pop counts with four auxiliary counts to flag outliers

- GQAC Expected Count
- GQAX Max Number of People
- Current Size
- Max Number of People

If the ratio between the reported pop count and the auxiliary pop count is determined to be an outlier, both the reported pop count and auxiliary count are removed from the models.

[Add a table showing how often we remove GQs from models]

Commented [JEZ(F14]: I know this is what we use, but consider editing to avoid confusion with hot deck methods. In econ we used 'imputation base' to refer to the cases that contribute to the ratios.

Commented [TLK(F15R14]: Good point. I agree with changing or droppiong the "donor pool" phrase. "Imputaiton base" avoids confusion with the hot deck, but it is a little bit confusing because it has the word "imputation" in it. I'll try to think of another phrase.
We could remove the text altogether, so it would be "contribute to the imputation."

Commented [PJC(F16]: Agree with your comments, and Tim's final suggestion.

Possible Methods

First, if a pop count is available from the NPC call operation and passes a quality review, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will use a combination of the following methods:

1. Ratio Imputation
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

Commented [ADK(F17]: Need to look at paradata as covariates as well on the models

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 5 shows that 8,600 of the unresolved GQ can be resolved by converting the GQAC expected count to the GQ pop count using the ratio adjustment.

Table 5: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For each GQ type, we will use the resolved cases with a GQAC expected count to calculate the ratio of the reported GQ Census Day count to the GQAC expected count. We will then use this ratio to convert the GQAC expected count of the unresolved GQs into a Census Day imputed count. For example, for an unresolved College GQ with a GQAC Expected Count, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. Conversion factors for the four variables under consideration are shown in Table 6. Table 9 and Table 11 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 6: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities				
Juvenile Facilities				
Nursing Facilities				
Hospitals				
College Housing				
Military				
Shelters				
Group Homes				
Other				
All GQs				

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for colleges and universities (GQTYPCUR=501).

First, we will adjust the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population, GQAC Max Number of People, and Greek Housing variables.

After adjusting the college-level total room capacity to account reference year and for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Once we calculate the college-level residual, we will then allocate the population counts among the GQs in the college without GQAC Expected Count.

Modeling

A third approach would be to impute the GQ pop counts from a Poisson regression model. The dependent variable will be log of the ratio of reported GQ pop count to GQ Current Max Size. Independent variables are

- GQ Type
-

See Table 3 for a description of the covariates. It is important to note that GQ type is a fixed-effect covariate in the model. Each model will contain the same set of covariates, with the exception of the college model, which will also include an indicator for Greek Housing.

Median Imputation

If sufficient auxiliary data is not available, we will impute the pop size with the median population size of the resolved GQs within GQ type and state. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and state. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median GQ pop size in the cell.

Evaluation of Imputed Values

We will evaluate the imputation methods using 10-fold cross validation. First we will remove the unresolved GQs from the universe since we don’t have a reported GQ pop for them. Second, we divide the remaining resolved GQs into 10 approximatley equal sized groups.

We will build and fit our models on nine of the groups and then impute responses for the remaining group. We will use all four methods to impute as many units in the “unresolved” group as possible. We will do this ten times, each time treating a different group as unresolved.

Then, for each group, we will calculate the difference between the reported GQ pop and each of the four imputed methods. For each group, we will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value. We will then average these statistics across the 10 groups.

Some methods may only work under certain conditions. For example, the IPEDS residual method will only work for colleges. The Poissoin regression will only work when all of the necessary covariates are filled. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Appendix

Table 7: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Hidiroglou-Berthelot (HB) Edits

The Hidiroglou-Berthelot (HB) edit detects outliers based on the ratio of two variables. In calculating the HB statistic, the ratio is transformed once to ensure that outliers are identified at both tails of the HB statistic's distribution, then transformed again to account for the size of the reporting unit. This results in identifying the records whose data exhibit the most unusual differences between the numerator and denominator variables as well as those that have more impact on the totals. These data are identified as requiring analyst review, suppression from the imputation donor pool, or imputation.

For our purposes in this project, the HB statistic was calculated as follows.

First, we calculated the ratio between the reported pop count and the auxiliary pop count for each GQ with positive counts for both values.

$$R_i = {}^{x_i}/_{y_i}$$

$$x_i \ = Reported\ Pop\ Count$$

$$y_i = Auxiliary\ Pop\ Count$$

We then transformed the ratios in order to detect outliers at both tails of the distribution. We calculated median ratios within each GQ type.

$$S_i = \begin{cases} 1 - \dfrac{R_{med}}{R_i} & 0 < \ R_i < \ R_{med} \\ \dfrac{R_i}{R_{med}} - 1 & R_i > R_{med} \end{cases}$$

$$R_{med} = median\ R_i$$

When then scaled the transformed ratios by GQ size to calculate the HB statistic.

$$E_i = \ S_i * \ \sqrt{\{\max{(x_i, y_i)}\}}$$

To detect outliers, we calculated the following values.

$$D_{Q1} = max\{E_{med} - E_{Q1}, |.05 * E_{med}|\}$$

$$D_{Q3} = max\{E_{Q3} - E_{med}, |.05 * E_{med}|\}$$

$$E_{med} = the\ median\ value\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q1} = the\ first\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q3} = the\ first\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

The outliers follow outside the following range.

$$\{E_{med} - c_j * D_{Q1}, E_{med} + c_j * D_{Q3}\}$$

$$c_j = parameter\ that\ controls\ the\ width\ of\ the\ acceptance\ interval$$

We set three C values for each GQ type. The C values determined the bounds for review, suppress, and impute flags. We conducted a manual review by plotting x_i and y_i values to set the bounds by GQ type. Note, this review is somewhat subjective, but imitates common practice for establishment surveys.

References

Hidiroglou, M.A., and Berthelot, J.-M. (1986). "Statistical Editing and Imputation for Periodic Business Surveys". Survey Methodology, 12, 73-83.

Table 8: GQAC Expected Count by Imputation Status

Commented [PJC(F18]: Will Tables 8 - 11 be adjusted when we determine the total set of unresolved cases, including the implauible cases with a response > 0?

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Table 9: GQAC Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Table 10: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Table 11: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

Group Quarters Imputation Methodology

Commented [JEZ(F1]: New version. Somehow I have locked myself out of version 3.

Table of Contents

Table of Tables

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

Commented [JEZ(F2]: Tables based on 12/18/20 data.

Commented [JEZ(F3R2]: Ryan will have a new file available 12/20/20. Once I get it I can re-run HB and recreate the tables in the doc.

A special telephone operation was conducted to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation that pass an initial quality review as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that open on Census Day, but vacant during the GQ Enumeration visit (which started in July 2020) require imputation.

In addition, we will impute a pop size for GQs that do not meet our quality edits. These GQs have a reported population that is not plausible and likely a error. For example, a facility might have reported their facilty population in one GQ. We employed the Hidiroglou-Berthelot (HB) editing process, commonly used in establishment surveys, to identify these GQs with an implausible reported population. Information about HB edits is included in the Appendix.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have an unplausible reported count. This universe is made up of GQs with a status of Occupied; Open on Census Day, but Vacant During Visit[1]; and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with implausible population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe. The other statuses of Vacant, Delete, and Nonresidential are considered out-of-scope for GQ Count Imputation. The GQ Count Imputaiton will only impute a positive population.

Table 1: GQ Universe

GQ Status	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Occupied GQ	177,000	17,000	3,100	197,000
Open on Census Day, Vacant During Visit	1,700	19,500	200	21,500
Refusal GQ	900	6,700	200	7,800
Vacant GQ	30,500	0	0	30,500
Delete GQ	7,600	0	0	7,600
Nonresidential GQ	2,500	0	0	2,500
Total	220,000	43,000	3,500	267,000

Commented [JEZ(F4]: Do we want to do any tables about the GEO review? Or will we just re-create this table after we incorporate the 'N's as resolved?

Commented [TLK(F5R4]: I think this table should be updated once the review is complete. Thus the N will be included in the resolved column.

[1] During GQ enumeration, the GQ was found to be vacant, but the contact at the GQ said the GQ was open on Census Day. This is different from the vacant GQs which were reported to be vacant on Census Day.

Some of the occupied GQs with a reported population were treated as unresolved because their census day population was implausible. The goal of the GQ Count Imputation is to determine a population count for all 43,000 occupied GQs with no reported population as well as the 3,500 occupied GQs with an implausible population count.

Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 7 in the Appendix has a full list of the GQ type codes.

Table 2: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

GQ Type	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Correctional Facilities*	13,000	2,800	250	16,000
Juvenile Facilities	6,100	1,800	90	8,000
Nursing Facilities*	24,500	3,200	550	28,500
Hospitals	1,900	800	70	2,800
College Housing*	29,000	5,500	1,300	36,000
Military*	3,000	1,900	70	5,000
Shelters	24,500	8,200	150	33,000
Group Homes	62,000	9,100	700	72,000
Other	16,000	9,700	300	26,000
Total	180,000	43,000	3,500	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Commented [PJC(F6]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Variables

Table 3 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, Administrative Records, and nursing home data from the Centers for Medicare & Medicaid Services (CMS). We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 3: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR
Number of All Beds	Total number of resident beds in the facility as reported by the provider.	CMS
Number of Occupied Beds	Total number of resident beds that are currently occupied as reported by the provider.	CMS

Commented [JEZ(F7]: From chat in EGG meeting: use 5-year ACS estimates?

Commented [JEZ(F8R7]: From James Christy: FWIW - when we visit a GQ for ACS, we ask for the total population of that GQ, then use that to sub-sample for selecting cases for interview. Reference date is when we visit the GQ. (Unlike Decennial which has a fixed reference date). I don't think that total pop count is part of what's published - but could be wrong.

Commented [TLK(F9R7]: Stuart Irby confirmed what James says. The Current GQ Size contains the size when conducting the listing. Current Surveys also updates the MAF in the same way as ACS.

Commented [JEZ(F10]: Add nursing home data from CMS. MEPS data?

An additional source available for college housing GQs is the Integrated Postsecondary Education Data System (IPEDS). These variables are available at the facility level but not for individual MAFIDs.

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

Commented [JEZ(F11]: We could include info in the appendix regarding matching.

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) "**capacity utilization**"—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Additional facility-level variables may become available as research continues.

Table 4: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Data Editing

The HB edits employed to detect implausible pop counts will also be used to determine which resolved GQs contribute to the donor pool for imputation. While the most extreme outliers are flagged for imputation, for less extreme outliers, we will accept the reported values, but keep those GQs from contributing to the imputation. We compared reported pop counts with four auxiliary counts to flag outliers

Commented [JEZ(F12]: Check tense.

Commented [JEZ(F13]: I know this is what we use, but consider editing to avoid confusion with hot deck methods. In econ we used 'imputation base' to refer to the cases that contribute to the ratios.

Commented [TLK(F14R13]: Good point. I agree with changing or droppiong the "donor pool" phrase. "Imputaiton base" avoids confusion with the hot deck, but it is a little bit confusing because it has the word "imputation" in it. I'll try to think of another phrase.
We could remove the text altogether, so it would be "contribute to the imputation."

- GQAC Expected Count
- GQAX Max Number of People
- Current GQ Size
- Max Number of People

If the ratio between the reported pop count and the auxiliary pop count is determined to be an outlier, both the reported pop count and auxiliary count are removed from the models. Note, the HB edit takes the GQ size into account, the ratios are transformed so that more importance is placed on a small deviation from the median ratio for a large GQ as opposed to a large deviation for a small GQ (Hidiroglou and Berthelot, 1986).

GQ Type	Suppressed from Models					
	GQAC Expected Count	GQAC Max Number of People	Current GQ Size	Max Number of People	Total Suppressed	Total Resolved
Correctional Facilities	N<15	N<15	90	90	150	13,000
Juvenile Facilities	20	30	40	60	90	6,100
Nursing Facilities	20	N<15	20	40	80	24,500
Hospitals	20	N<15	20	30	50	1,900
College Housing	250	N<15	50	80	400	29,000
Military	N<15	20	20	N<15	50	3,000
Shelters	30	N<15	30	100	150	24,500
Group Homes	60	N<15	150	30	200	62,000
Other	30	20	30	70	150	16,000
Total	450	150	450	500	1,300	180,000

In addition to outliers flagged for imputation or to be suppressed from our models, a third set of less extreme outliers will be identified with a flag for review. These flags may help to prioritize subject-matter review of final GQ counts.

Possible Methods

First, if a pop count is available from the NPC call operation and passes a quality review, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will use a combination of the following methods:

1. Ratio Imputation
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

Commented [ADK(F15]: Need to look at paradata as covariates as well on the models

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 5 shows that 8,600 of the unresolved GQ can be resolved by converting the GQAC expected count to the GQ pop count using the ratio adjustment.

Table 5: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For each GQ type, we will use the resolved cases with a GQAC expected count to calculate the ratio of the reported GQ Census Day count to the GQAC expected count. We will then use this ratio to convert the GQAC expected count of the unresolved GQs into a Census Day imputed count. For example, for an unresolved College GQ with a GQAC Expected Count, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. Conversion factors for the four variables under consideration are shown in Table 6. Table 9 and Table 11 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 6: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities				
Juvenile Facilities				
Nursing Facilities				
Hospitals				
College Housing				
Military				
Shelters				
Group Homes				
Other				
All GQs				

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for colleges and universities (GQTYPCUR=501).

First, we will adjust the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population, GQAC Max Number of People, and Greek Housing variables.

After adjusting the college-level total room capacity to account reference year and for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Once we calculate the college-level residual, we will then allocate the population counts among the GQs in the college without GQAC Expected Count.

Modeling

A third approach would be to impute the GQ pop counts from a Poisson regression model. The dependent variable will be log of the ratio of reported GQ pop count to GQ Current Max Size. Independent variables are

- GQ Type
-

See Table 3 for a description of the covariates. It is important to note that GQ type is a fixed-effect covariate in the model. Each model will contain the same set of covariates, with the exception of the college model, which will also include an indicator for Greek Housing.

Median Imputation

If sufficient auxiliary data is not available, we will impute the pop size with the median population size of the resolved GQs within GQ type and state. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and state. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median GQ pop size in the cell.

Evaluation of Imputed Values

We will evaluate the imputation methods using 10-fold cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we divide the remaining resolved GQs into 10 approximatley equal sized groups.

We will build and fit our models on nine of the groups and then impute responses for the remaining group. We will use all four methods to impute as many units in the "unresolved" group as possible. We will do this ten times, each time treating a different group as unresolved.

Then, for each group, we will calculate the difference between the reported GQ pop and each of the four imputed methods. For each group, we will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value. We will then average these statistics across the 10 groups.

Some methods may only work under certain conditions. For example, the IPEDS residual method will only work for colleges. The Poissoin regression will only work when all of the necessary covariates are filled. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Appendix

Table 7: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Hidiroglou-Berthelot (HB) Edits

The Hidiroglou-Berthelot (HB) edit detects outliers based on the ratio of two variables. In calculating the HB statistic, the ratio is transformed once to ensure that outliers are identified at both tails of the HB statistic's distribution, then transformed again to account for the size of the reporting unit. This results in identifying the records whose data exhibit the most unusual differences between the numerator and denominator variables as well as those that have more impact on the totals. These data are identified as requiring analyst review, suppression from the imputation donor pool, or imputation.

For our purposes in this project, the HB statistic was calculated as follows.

First, we calculated the ratio between the reported pop count and the auxiliary pop count for each GQ with positive counts for both values.

$$R_i = {}^{x_i}/_{y_i}$$

$$x_i \ = Reported\ Pop\ Count$$

$$y_i = Auxiliary\ Pop\ Count$$

We then transformed the ratios in order to detect outliers at both tails of the distribution. We calculated median ratios within each GQ type.

$$S_i = \begin{cases} 1 - \frac{R_{med}}{R_i} & 0 < \ R_i < \ R_{med} \\ \frac{R_i}{R_{med}} - 1 & R_i > R_{med} \end{cases}$$

$$R_{med} = median\ R_i$$

When then scaled the transformed ratios by GQ size to calculate the HB statistic.

$$E_i = \ S_i * \ \sqrt{\{\max{(x_i, y_i)}\}}$$

To detect outliers, we calculated the following values.

$$D_{Q1} = max\{E_{med} - E_{Q1}, |.05 * E_{med}|\}$$

$$D_{Q3} = max\{E_{Q3} - E_{med}, |.05 * E_{med}|\}$$

$$E_{med} = the\ median\ value\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q1} = the\ first\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q3} = the\ third\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

The outliers follow outside the following range.

$$\{E_{med} - c_j * D_{Q1}, E_{med} + c_j * D_{Q3}\}$$

$$c_j = parameter\ that\ controls\ the\ width\ of\ the\ acceptance\ interval$$

We set three C values for each GQ type. The C values determined the bounds for review, suppress, and impute flags. We conducted a manual review by plotting x_i and y_i values to set the bounds by GQ type. Note, this review is somewhat subjective, but imitates common practice for establishment surveys.

Matching to IPEDS

Matching to CMS Nursing Home Data

References

Hidiroglou, M.A., and Berthelot, J.-M. (1986). "Statistical Editing and Imputation for Periodic Business Surveys". Survey Methodology, 12, 73-83.

Table 8: GQAC Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Table 9: GQAC Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Table 10: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Table 11: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

Group Quarters Imputation Methodology

Commented [JEZ(F1]: New version. Somehow I have locked myself out of version 3.

Table of Contents

Table of Tables

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

Commented [JEZ(F2]: Tables based on 12/18/20 data.

Commented [JEZ(F3R2]: Ryan will have a new file available 12/20/20. Once I get it I can re-run HB and recreate the tables in the doc.

Commented [JMA(F4]: I suggest adding a table with the sources of the data used to classify the MAFID as "occupied group quarters." Be clear about which 2020 Census operation so classified each one. Summarize here.

Commented [JMA(F5]: Suggest adding "Large GQs are often the only addresses in their tabulation census block. Consequently, information suggesting that such a GQ has very low population, or zero, will be evident in the PL 94-171 redistricting data summary file."

A special telephone operation was conducted to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation that pass an initial quality review as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that were open on Census Day, but were vacant during the GQ Enumeration visit (which ~~started in~~occured between late July ~~2020~~and mid-October 2020) require imputation.

In addition, we will impute a population size for GQs that do not meet our quality edits, as implemented in the DRF1 review process. These GQs have a reported population that is not plausible and likely ~~a~~ ~~error~~erroneous. The criteria for implausibility differ by GQ type. They are based on historical information maintained in the MAF database and other sources available to the DSSD, POP and SEHSD reviewers. They were consistently applied to all GQ reports. For example, a facility, which is a collection of GQs operated by a single reporting organization, might have reported the entire~~ir~~ facilty population in one GQ MAFID. We employed the Hidiroglou-Berthelot (HB) editing process, commonly used in establishment surveys, to identify these GQs with an implausible reported population. Information about HB edits is included in the Appendix.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that ~~are expected to be~~were enumerated in a status related to occupied, but (1) do not have a reported count, or (2) have an ~~unplausible~~ implausible reported count. This universe is made up of GQs with a status of Occupied; Open on Census Day, but Vacant During Visit[1]; and Refusals. ~~Altogether, w~~We ~~call~~ consider these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with implausible population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe. The other statuses of Vacant, Delete, and Nonresidential are considered out-of-scope for GQ Count Imputation. The GQ Count Imputaiton will only impute a positive population.

Table 1: GQ Universe

GQ Status	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Occupied GQ	177,000	17,000	3,100	197,000

Commented [JEZ(F6]: Do we want to do any tables about the GEO review? Or will we just re-create this table after we incorporate the 'N's as resolved?

Commented [TLK(F7R6]: I think this table should be updated once the review is complete. Thus the N will be included in the resolved column.

[1] During GQ enumeration, the GQ was found to be vacant, but the contact at the GQ said the GQ was open on Census Day. This is different from the vacant GQs which were reported to be vacant on Census Day.

Open on Census Day, Vacant During Visit	1,700	19,500	200	21,500
Refusal GQ	900	6,700	200	7,800
Vacant GQ	30,500	0	0	30,500
Delete GQ	7,600	0	0	7,600
Nonresidential GQ	2,500	0	0	2,500
Total	220,000	43,000	3,500	267,000

Some of the occupied GQs with a reported population were treated as unresolved because their census day population was implausible. The goal of the GQ Count Imputation is to determine a population count for all 43,000 occupied GQs with no reported population as well as the 3,500 occupied GQs with an implausible population count.

Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 7 in the Appendix has a full list of the GQ type codes.

Table 2: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

GQ Type	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Correctional Facilities*	13,000	2,800	250	16,000
Juvenile Facilities	6,100	1,800	90	8,000
Nursing Facilities*	24,500	3,200	550	28,500
Hospitals	1,900	800	70	2,800
College Housing*	29,000	5,500	1,300	36,000
Military*	3,000	1,900	70	5,000
Shelters	24,500	8,200	150	33,000
Group Homes	62,000	9,100	700	72,000
Other	16,000	9,700	300	26,000
Total	180,000	43,000	3,500	227,000

*denotes GQ Type is included in NPC calling operation

In order to avoid duplicated persons in the GQ MAFIDs that will receive imputed population counts, responses found among records identified in the unduplication as persons properly residing in one of the unresolved GQ MAFIDs will be assigned to that MAFID and subtracted from the imputed population. This avoids double counting such people.

Imputation Methods

Variables

Table 3 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, Administrative Records, and nursing home data from the Centers for Medicare & Medicaid Services (CMS). We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Commented [JMA(F8]: We should ensure that some version of this sentence is implemented. I think the first part automatic in unduplication from DRF1 to DRF2. Not sure about the subtraction part. It will be important to actively document our unduplication efforts here. The HB algorithm also addresses this, but it is not transparent to a lay reader.

Commented [PJC(F9]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Commented [JMA(F10R9]: Agree with Pat.

Table 3: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR
Number of All Beds	Total number of resident beds in the facility as reported by the provider.	CMS
Number of Occupied Beds	Total number of resident beds that are currently occupied as reported by the provider.	CMS

Commented [JEZ(F11]: From chat in EGG meeting: use 5-year ACS estimates?

Commented [JEZ(F12R11]: From James Christy: FWIW - when we visit a GQ for ACS, we ask for the total population of that GQ, then use that to sub-sample for selecting cases for interview. Reference date is when we visit the GQ. (Unlike Decennial which has a fixed reference date). I don't think that total pop count is part of what's published - but could be wrong.

Commented [TLK(F13R11]: Stuart Irby confirmed what James says. The Current GQ Size contains the size when conducting the listing. Current Surveys also updates the MAF in the same way as ACS.

Commented [JEZ(F14]: Add nursing home data from CMS. MEPS data?

An additional source available for college housing GQs is the Integrated Postsecondary Education Data System (IPEDS). These variables are available at the facility level but not for individual MAFIDs.

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

Commented [JEZ(F15]: We could include info in the appendix regarding matching.

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) "**capacity utilization**"—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Additional facility-level variables may become available as research continues.

Table 4: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Data Editing

The HB edits employed to detect implausible population counts will also be used to determine which resolved GQs contribute to the ~~donor~~ imputation base~~pool~~ ~~for imputation~~. While the most extreme outliers are flagged for imputation, for less extreme outliers, we will accept the reported values, but keep those GQs from contributing to the statistical estimation that produces the imputation. We compared reported pop counts with four auxiliary counts to flag outliers

Commented [JEZ(F16]: I know this is what we use, but consider editing to avoid confusion with hot deck methods. In econ we used 'imputation base' to refer to the cases that contribute to the ratios.

Commented [TLK(F17R16]: Good point. I agree with changing or droppiong the "donor pool" phrase. "Imputaiton base" avoids confusion with the hot deck, but it is a little bit confusing because it has the word "imputation" in it. I'll try to think of another phrase.
We could remove the text altogether, so it would be "contribute to the imputation."

Commented [JMA(F18R16]: Absolutely. This is not a hot deck, but it is a legitimate and well tested (albeit in other contexts) statistical imputation model. Use the standard Econ terms.

- GQAC Expected Count
- GQAX Max Number of People
- Current GQ Size
- Max Number of People

If the ratio between the reported population count and the auxiliary population count is determined to be an outlier, both the reported ~~pop~~ count and ~~auxiliary~~ count are removed from the estimated models. Note, the HB edit takes the GQ size into account, the ratios are transformed so that more importance is placed on a small deviation from the median ratio for a large GQ as opposed to a large deviation for a small GQ (Hidiroglou and Berthelot, 1986).

[Add a table showing how often we remove GQs from imputation base for the statistical models]

In addition to outliers flagged for imputation or to be ~~suppressed~~ eliminated from the imputation base ~~from~~ for our models, a third set of less extreme outliers will be identified with a flag for review. These flags may help to prioritize subject-matter review of final GQ counts.

~~Possible~~ Candidate Methods

First, if a population count is available from the NPC call operation and passes ~~a~~ standard DRF1 quality review, we will use that ~~pop~~ count as a response and not impute ~~a pop size~~the population for that MAFID.

The GQ count imputation will use a combination of the following methods:

1. Ratio Imputation
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

Commented [ADK(F19]: Need to look at paradata as covariates as well on the models

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ population count, as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error relative to the Census Day target of April 1 than other methods. Our research on GQs that reported during GQE should provide information on this ~~presumption~~assumption, and on functions of the expected GQ population count that produce more accurate imputation.

Table 5 shows that 8,600 of the unresolved GQ can be resolved by converting the GQAC expected count to the imputed GQ population count using the ratio adjustment.

Table 5: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For each GQ type, we will use the resolved cases with a GQAC expected count to calculate the ratio of the reported GQ Census Day count to the GQAC expected count. We will then use this ratio to convert the GQAC expected count of the unresolved GQs into a Census Day imputed count. For example, for an unresolved College GQ with a GQAC Expected Count, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. Conversion factors for the four variables under consideration are shown in Table 6. Table 9 and Table 11 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 6: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities				
Juvenile Facilities				
Nursing Facilities				
Hospitals				
College Housing				
Military				
Shelters				
Group Homes				
Other				
All GQs				

Adjusted Residual from Facility-level Total for College Housing

~~A~~ The second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for colleges and universities (GQTYPCUR=501).

First, we will adjust the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population, GQAC Max Number of People, and Greek Housing variables.

After adjusting the college-level total room capacity to account for reference year and ~~for~~ capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Population\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Commented [JMA(F20]: We should call out that the Reported GQ Population Count contains the persons who were properly placed in the GQ during unduplication.

Once we calculate the college-level residual, we will then allocate the population counts among the GQs in the college without GQAC Expected Count.

Modeling

~~A~~ The third approach would be to impute the GQ population counts from a Poisson regression model. The dependent variable ~~will~~ would be natural logarithm of the ratio of reported GQ population count to GQ Current Max Size. Independent variables are

- GQ Type
-

See Table 3 for a description of the covariates. It is important to note that GQ type is a fixed-effect covariate in the model. Each model will contain the same set of covariates, with the exception of the college model, which will also include an indicator for Greek Housing.

Median Imputation

If sufficient auxiliary data ~~is~~ are not available, we will impute the population size with the median population size of the resolved GQs within GQ type and state. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and state. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median GQ pop size in the cell.

Evaluation of Imputed Values

We will evaluate the imputation methods using 10-fold cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we divide the remaining resolved GQs into 10 approximatley equal sized groups.

Commented [JMA(F21]: I see someone was listening. I think this is a good choice. If it turns out to be to computationally intensive, switch to 5-fold.

We will build and fit our models on nine of the groups and then impute responses for the remaining group, repeating this process 10 times, so that an out-of-sample forecast error is available for each observation in the imputation base used for estimation. We will ~~use~~ evaluate all four methods, ~~to~~

imput~~e~~ing as many units in the “unresolved” group as possible. ~~We will do this ten times, each time treating a different group as unresolved.~~

Then, for each group, we will calculate the difference between the reported GQ population and ~~each~~ the k-fold predicted value for of the four imputed methods. This generates the out-of-sample forecast errors. For each group, we will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value. We will then average these statistics across the 10 groups.

Commented [JMA(F22]: This is not the correct procedure for k-fold cross validation. If N is the total number of observations, then for each i in N, you have exactly one k-fold forecast error for each method. Then number of groups (k=10) is no longer relevant. You compute the statistics for each of the four methods from the N out-of-sample forecast errors. The method described in the text is only correct if the aggregate error measure is linear in the out-of-sample forecasts. Details of for other error measures can be found in the Rodriquez et al (2010) article I put in the same directory as these notes.

Some methods may only work under certain conditions. For example, the IPEDS residual method will only work for colleges. The Poisso~~i~~n regression will only work when all of the necessary covariates are filled. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Appendix

Table 7: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Hidiroglou-Berthelot (HB) Edits

The Hidiroglou-Berthelot (HB) edit detects outliers based on the ratio of two variables. In calculating the HB statistic, the ratio is transformed once to ensure that outliers are identified at both tails of the HB statistic's distribution, then transformed again to account for the size of the reporting unit. This results in identifying the records whose data exhibit the most unusual differences between the numerator and denominator variables as well as those that have more impact on the totals. These data are identified as requiring analyst review, suppression from the imputation donor pool, or imputation.

For our purposes in this project, the HB statistic was calculated as follows.

First, we calculated the ratio between the reported pop count and the auxiliary pop count for each GQ with positive counts for both values.

$$R_i = {}^{x_i}/_{y_i}$$

$$x_i \ = Reported\ Population\ Count$$

$$y_i = Auxiliary\ Population\ Count$$

We then transformed the ratios in order to detect outliers at both tails of the distribution. We calculated median ratios within each GQ type.

$$S_i = \begin{cases} 1 - \dfrac{R_{med}}{R_i} & 0 < \ R_i < \ R_{med} \\ \dfrac{R_i}{R_{med}} - 1 & R_i > R_{med} \end{cases}$$

$$R_{med} = median\ R_i$$

When then scaled the transformed ratios by GQ size to calculate the HB statistic.

$$E_i = \ S_i * \ \sqrt{\{\max{(x_i, y_i)}\}}$$

To detect outliers, we calculated the following values.

$$D_{Q1} = max\{E_{med} - E_{Q1}, |.05 * E_{med}|\}$$

$$D_{Q3} = max\{E_{Q3} - E_{med}, |.05 * E_{med}|\}$$

$$E_{med} = the\ median\ value\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q1} = the\ first\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q3} = the\ first\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

The outliers follow outside the following range.

$$\{E_{med} - c_j * D_{Q1}, E_{med} + c_j * D_{Q3}\}$$

$$c_j = parameter\ that\ controls\ the\ width\ of\ the\ acceptance\ interval$$

We set three C values for each GQ type. The C values determined the bounds for review, suppress, and impute flags. We conducted a manual review by plotting x_i and y_i values to set the bounds by GQ type. Note, this review is somewhat subjective, but imitates common practice for establishment surveys.

Matching to IPEDS

Matching to CMS Nursing Home Data

References

Hidiroglou, M.A., and Berthelot, J.-M. (1986). ”Statistical Editing and Imputation for Periodic Business Surveys”. Survey Methodology, 12, 73-83.

Table 8: GQAC Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Table 9: GQAC Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Table 10: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Table 11: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

Disclosure Prohibited. Title 13 U.S. Code

Group Quarters Imputation Methodology

Commented [JEZ(F1]: New version. Somehow I have locked myself out of version 3.

Commented [JEZ(F2R1]: Version 3 2 combines comments from John Abowd and Pat.

Table of Contents

Table of Tables

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, especially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic. Large GQs are often the only addresses in their tabulation census block. Consequently, information suggesting that such a GQ has very low population, or zero, will be evident in the PL 94-171 redistricting data summary file.

Commented [JEZ(F3]: Tables based on 12/18/20 data.

Commented [JEZ(F4R3]: Ryan will have a new file available 12/20/20. Once I get it I can re-run HB and recreate the tables in the doc.

Commented [JEZ(F5R3]: Research will use data as of 12/13/20. GEO counts will be used in production.

Commented [JMA(F6]: I suggest adding a table with the sources of the data used to classify the MAFID as "occupied group quarters." Be clear about which 2020 Census operation so classified each one. Summarize here.

Commented [JEZ(F7R6]: Need to ask Ryan and Debbie.

Commented [JMA(F8]: Suggest adding "Large GQs are often the only addresses in their tabulation census block. Consequently, information suggesting that such a GQ has very low population, or zero, will be evident in the PL 94-171 redistricting data summary file."

A special telephone operation was conducted to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation that pass an initial quality review as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that were open on Census Day, but were vacant during the GQ Enumeration visit (which occurred between late July and mid-October 2020) require imputation.

In addition, we will impute a population size for GQs that do not meet our quality edits, as implemented in the DRF1 review process. These GQs have a reported population that is not plausible and likely erroneous. The criteria for implausibility differ by GQ type. They are based on historical information maintained in the MAF database and other sources available to the DSSD, POP and SEHSD reviewers. They will be consistently applied to all GQ reports.~~.~~ For example, a facility, which is a collection of GQs operated by a single reporting organization, might have reported the entire facility population in one GQ MAFID. We will employ the Hidiroglou-Berthelot (HB) editing process, commonly used in establishment surveys, to identify these GQs with an implausible reported population. Information about HB edits is included in the Appendix.

This document details a joint research effort by staff in DSSD and CES to determine a method for GQ Count Imputation. A specification will be written to detail the final method that is implemented in production. The data in this document represent the GQ Universe as of December 13, 2020. This universe formed the basis of our research into possible imputation methods. The universe in production will change as a result of the NPC calling operation and subsequent review by staff in GEO. Final imputation results will be provided to the POP division for subject-matter-review.

Commented [JEZ(F9]: Tim, check that all this is accurate. I think we need to pin down the universe. It's hard to write up what we're doing when it keeps changing. I'm thinking this document will be just the research universe and recommendations and if necessary, after we finish, we can write up a very short memo with final results. That's what makes sense to me. I don't want to keep mixing the imputation research with what is happening in production. Also, I don't have the expertise to write about what Ryan is doing.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that ~~are expected to be~~were enumerated in a status related to occupied, but (1) do not have a reported count, or (2) have an implausible reported count. This universe is made up of GQs with a status of Occupied; Open on Census Day, but Vacant During Visit[1]; and Refusals. We consider these GQs unresolved and will impute a count for them. Some of the occupied GQs with a reported population were treated as unresolved because their census day population was implausible. The goal of the GQ Count Imputation is to determine a

[1] During GQ enumeration, the GQ was found to be vacant, but the contact at the GQ said the GQ was open on Census Day. This is different from the vacant GQs which were reported to be vacant on Census Day.

population count for all 43 000 occupied GQs with no reported population as well as the 3 500 occupied GQs with an implausible population count.

Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. ~~In Table 1, the GQs with implausible population count are included in the Census Day Pop column.~~ The first three rows represent the occupied GQ universe. The other statuses of Vacant, Delete, and Nonresidential are considered out-of-scope for GQ Count Imputation. The GQ Count Imputation will only impute a positive population count.

Table 1: GQ Universe as of December 13, 2020

GQ Status	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Occupied GQ	177,000	17,000	3,100	197,000
Open on Census Day, Vacant During Visit	1,700	19,500	200	21,500
Refusal GQ	900	6,700	200	7,800
Vacant GQ	30,500	0	0	30,500
Delete GQ	7,600	0	0	7,600
Nonresidential GQ	2,500	0	0	2,500
Total	220,000	43,000	3,500	267,000

~~Some of the occupied GQs with a reported population were treated as unresolved because their census day population was implausible. The goal of the GQ Count Imputation is to determine a population count for all 43,000 occupied GQs with no reported population as well as the 3,500 occupied GQs with an implausible population count.~~

Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 7 in the Appendix has a full list of the GQ type codes.

Table 2: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

GQ Type	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Correctional Facilities*	13,000	2,800	250	16,000
Juvenile Facilities	6,100	1,800	90	8,000
Nursing Facilities*	24,500	3,200	550	28,500
Hospitals	1,900	800	70	2,800
College Housing*	29,000	5,500	1,300	36,000
Military*	3,000	1,900	70	5,000
Shelters	24,500	8,200	150	33,000
Group Homes	62,000	9,100	700	72,000
Other	16,000	9,700	300	26,000
Total	180,000	43,000	3,500	227,000

*denotes GQ Type is included in NPC calling operation

In order to avoid duplicated persons in the GQ MAFIDs that will receive imputed population counts, responses found among records identified in the unduplication as persons properly residing in one of the unresolved GQ MAFIDs will be assigned to that MAFID and subtracted from the imputed population. This avoids double counting such people.

Commented [JEZ(F10]: Do we want to do any tables about the GEO review? Or will we just re-create this table after we incorporate the 'N's as resolved?

Commented [TLK(F11R10]: I think this table should be updated once the review is complete. Thus the N will be included in the resolved column.

Commented [PJC(F12]: I agree. And somewhere we can describe briefly the calling operation, the data collected in it, and GEO's review.

Commented [JEZ(F13R12]: See my comment above.

Commented [PJC(F14]: We should insert "occupied GQs" or something similar (occupied or suspected occupied) in the table title somewhere.

Commented [JMA(F15]: We should ensure that some version of this sentence is implemented. I think the first part automatic in unduplication from DRF1 to DRF2. Not sure about the subtraction part. It will be important to actively document our unduplication efforts here. The HB algorithm also addresses this, but it is not transparent to a lay reader.

Commented [JEZ(F16R15]: Need to ask Ryan about this.

Imputation Methods

Variables

Table 3 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, Administrative Records, and nursing home data from the Centers for Medicare & Medicaid Services (CMS). We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 3: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR
Number of All Beds	Total number of resident beds in the facility as reported by the provider.	CMS
Number of Occupied Beds	Total number of resident beds that are currently occupied as reported by the provider.	CMS

An additional source available for college housing GQs is the Integrated Postsecondary Education Data System (IPEDS). These variables are available at the facility level but not for individual MAFIDs.

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

(1) **reference year**—our latest IPEDS data is for reference year 2019;

Commented [PJC(F17]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Added: As our method has developed, and we're embracing the median imputation option when no data are available, it appears that this set will be empty.

Commented [JMA(F18R17]: Agree with Pat.

Commented [JEZ(F19]: From chat in EGG meeting: use 5-year ACS estimates?

Commented [JEZ(F20R19]: From James Christy: FWIW - when we visit a GQ for ACS, we ask for the total population of that GQ, then use that to sub-sample for selecting cases for interview. Reference date is when we visit the GQ. (Unlike Decennial which has a fixed reference date). I don't think that total pop count is part of what's published - but could be wrong.

Commented [TLK(F21R19]: Stuart Irby confirmed what James says. The Current GQ Size contains the size when conducting the listing. Current Surveys also updates the MAF in the same way as ACS.

Commented [JEZ(F22]: Add nursing home data from CMS. MEPS data?

Commented [JEZ(F23]: Have we dropped some of these? I think from what Andy showed today he was just using the 4 counts in the poisson? Maybe I heard that wrong.

Commented [JEZ(F24]: We could include info in the appendix regarding matching.

(2) **"capacity utilization"**—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

~~Additional facility-level variables may become available as research continues.~~

Table 4: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Data Editing

The HB edits employed to detect implausible pop counts can also be used to determine which resolved GQs contribute to the imputation. While the most extreme outliers are flagged for imputation, for less extreme outliers, we will accept the reported values, but keep those GQs from contributing to the imputation. We compared reported pop counts with four auxiliary counts to flag outliers

- GQAC Expected Count
- GQAX Max Number of People
- Current GQ Size
- Max Number of People

If the ratio between the reported pop count and the auxiliary pop count is determined to be an outlier, both the reported pop count and auxiliary count are removed from the models. Note, the HB edit takes the GQ size into account, the ratios are transformed so that more importance is placed on a small deviation from the median ratio for a large GQ as opposed to a large deviation for a small GQ (Hidiroglou and Berthelot, 1986). Table 5 shows counts of GQs that were suppressed from our imputation models for our research. The same GQ could be flagged for suppression by more than one outlying ratio, therefore the total number of suppressed GQs is not equal to the sum of the flags for each ratio.

Table 5: Counts of GQs suppressed from imputation models by GQ Type

GQ Type	Suppressed from Models					
	GQAC Expected Count	GQAC Max Number of People	Current GQ Size	Max Number of People	Total Suppressed	Total Resolved
Correctional Facilities	N<15	N<15	90	90	150	13,000
Juvenile Facilities	20	30	40	60	90	6,100
Nursing Facilities	20	N<15	20	40	80	24,500
Hospitals	20	N<15	20	30	50	1,900
College Housing	250	N<15	50	80	400	29,000
Military	N<15	20	20	N<15	50	3,000
Shelters	30	N<15	30	100	150	24,500
Group Homes	60	N<15	150	30	200	62,000
Other	30	20	30	70	150	16,000
Total	450	150	450	500	1,300	180,000

In addition to outliers flagged for imputation or to be suppressed from our models, a third set of less extreme outliers will be identified with a flag for review. These flags may help to prioritize subject-matter review of final GQ counts.

Candidate Methods

~~First, if a pop count is available from the NPC call operation and passes a quality review, we will use that pop count as a response and not impute a pop size.~~

~~The GQ count imputation will use a combination of the~~For GQ Count Imputation we evaluated the following methods:

1. Ratio Imputation
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. ~~Median~~ Percentile Imputation

Commented [ADK(F25]: Need to look at paradata as covariates as well on the models

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we ~~will~~ can use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. ~~Our research on GQs that reported during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.~~

Table 5 shows that ~~8,600~~ 11,000 of the unresolved GQs included in our research could ~~can~~ be resolved by converting the GQAC expected count to the GQ pop count using the ratio adjustment.

Commented [JEZ(F26]: Updated with the latest.

Commented [JEZ(F27R26]: We should think about what to do in production when we have a flag on the GP/Exp count ratio - we don't know which value is 'wrong'. We are trying to impute GP, so we could use expected count, but we don't want to do that if the expected count is what is off in the ratio. For the truth deck and for the donor pool, I think it's fine to throw out both but for production we need to figure out when we should accept GP or expected count (same applies to the other vars). For now, no rule has been applied to this table (i.e. ID could be unresolved because of an I flag on the GP/expected count ratio and still have expected count populated in this table).

Table 6: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	91,000	11,000	192,000
Not Populated	89,000	35,500	125,000

Total	180,000	46,500	227,000

For each detailed GQ type (see Table 8) within each state , we ~~will use~~used the resolved cases with a GQAC expected count to calculate the ratio of the reported GQ Census Day count to the GQAC expected count. We ~~will~~ then used this ratio to convert the GQAC expected count of the unresolved GQs into a Census Day imputed count. For example, for an unresolved College GQ with a GQAC Expected Count, the following equation would be applied:

$$Imputed\ Population\ Count = \ GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We ~~will construct~~constructed ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. ~~Conversion factors~~Ratios for the four variables ~~under consideration~~ by GQ Type are shown in Table 6. The ratios presented here were not used directly – we used the more detailed GQ type and state to calculate the ratios in the imputation. Table 9 ~~and~~ through Table 11 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 7: Factors to convert Auxiliary Variables to GQ Population

Commented [JEZ(F28]: Need to add a sentence about Andy using state-level ratios. These aren't exactly what is used.

GQ Type	Ratio of ~~Good People~~Reported Count to GQAC Expected Count	Ratio ~~of Good People~~of Reported Count to GQAC Max Number of People	Ratio of ~~Good People~~ Reported Count to Current GQ Size	Ratio of ~~Good People~~ Reported Count to Max Number of People
Correctional Facilities	0.7389	0 6613	0.8960	0.7328
Juvenile Facilities	0.7363	0.5870	0.8713	0.6636
Nursing Facilities	0.8710	0.7482	0.7925	0.5916
Hospitals	0.7925	0 6928	0.9360	0.7463
College Housing	0.9056	0.8069	0.8986	0.6881
Military	0.7540	0 6822	0.9249	0.8118
Shelters	0.6353	0.6115	0.8652	0.5490
Group Homes	0.8816	0.7792	0.6502	0.6703
Other	0.8453	0.6137	0 9216	0.7851
All GQs	0.8480	0.7342	0.8960	0.7320

Commented [JEZ(F29]: Removed all IDs for which ANY flags are 'S' or 'I'. This lines up with what Andy is using in the truth deck. In production, we probably will want to only exclude for certain ratios (i e.if GP/Exp count looks okay, keep in that ratio but GP/Max count is flagged, exclude from that ratio)

Commented [JEZ(F30]: Do we still want the Greek break-out? Need to ask Andy if he is using it.

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for colleges and universities (GQTYPCUR=501).

First, we ~~will~~ adjusted the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population, GQAC Max Number of People, and Greek Housing variables.

After adjusting the college-level total room capacity to account reference year and for capacity utilization, we ~~will~~ calculated the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Once we calculated the college-level residual, we ~~will~~ then allocated the population counts among the GQs in the college without GQAC Expected Count.

Modeling

A third approach would be to impute the GQ population counts from a Poisson regression model. The dependent variable will be log of the ratio of reported GQ pop count to GQ Current Max Size. Independent variables are

- GQ Type
- GQAC Expected Count
- GQAC Max Number of People
- Current GQ Size

See Table 3 for a description of the covariates. It is important to note that GQ type is a fixed-effect covariate in the model. Each model will contain the same set of covariates, ~~with the exception of~~except for the college model, which will also include an indicator for Greek Housing.

~~Median~~ Percentile Imputation

If sufficient auxiliary data is not available, we will impute the population count with the median population count of the resolved GQs within detailed GQ type and state. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and state. Then, we will calculate the median or other percentile of the GQ population count and impute the unresolved GQs with the median or other percentile of the GQ population count in the cell. We will determine the percentile to use based on the value that minimizes the imputation bias in our evaluation.

Evaluation of Imputed Values

We ~~will~~ evaluated the imputation methods using 10-fold cross validation. First, we ~~will~~ removed the unresolved GQs from the universe since we don't have a reported GQ population count for them. Next, we removed GQs with a count that was implausible based on our edits. We also removed any GQs that had any of the four flags set to suppress the reported population count from the imputation models. This ensured that extreme outliers would not influence our evaluation. ~~Second, we~~We then divided the remaining resolved GQs into 10 approximately equal sized groups.

Commented [JEZ(F31]: Need to add info about Andy's truth deck construction.

We ~~will build~~ built and fit our models on nine of the groups and then imputed responses for the remaining group. We ~~will~~ used all four methods to impute as many units in the “unresolved” group as possible. We ~~will do~~did this ten times, each time treating a different group as unresolved.

Then, for each group, we ~~will~~ calculated the difference between the reported GQ population count and the imputed value using each of the four imputed methods. For each group, we ~~will~~ summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We ~~will~~ also produce these metrics for the ratio of the imputed value and the reported value. We ~~will~~ then average these statistics across the 10 groups.

Some methods may only work under certain conditions. For example, the IPEDS residual method will only work for colleges. The Poisson regression will only work when all of the necessary covariates are filled. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Research Results

Commented [JEZ(F32]: Need to add some table shells. Even if they’re not filled in on Wednesday we can get some direction.

Disclosure Prohibited. Title 13 U.S. Code

Recommendation

Appendix

Table 8: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers

CODE	VALUE
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Hidiroglou-Berthelot (HB) Edits

The Hidiroglou-Berthelot (HB) edit detects outliers based on the ratio of two variables. In calculating the HB statistic, the ratio is transformed once to ensure that outliers are identified at both tails of the HB statistic's distribution, then transformed again to account for the size of the reporting unit. This results in identifying the records whose data exhibit the most unusual differences between the numerator and denominator variables as well as those that have more impact on the totals. These data are identified as requiring analyst review, suppression from the imputation donor pool, or imputation.

For our purposes in this project, the HB statistic was calculated as follows.

First, we calculated the ratio between the reported pop count and the auxiliary pop count for each GQ with positive counts for both values.

$$R_i = {}^{x_i}/y_i$$

$$x_i \ = Reported\ Pop\ Count$$

$$y_i = Auxiliary\ Pop\ Count$$

We then transformed the ratios in order to detect outliers at both tails of the distribution. We calculated median ratios within each GQ type.

$$S_i = \begin{cases} 1 - \dfrac{R_{med}}{R_i} & 0 < \ R_i < \ R_{med} \\ \dfrac{R_i}{R_{med}} - 1 & R_i > R_{med} \end{cases}$$

$$R_{med} \ = median\ R_i$$

When then scaled the transformed ratios by GQ size to calculate the HB statistic.

$$E_i \ = \ S_i * \ \sqrt{\{\max{(x_i, y_i)}\}}$$

To detect outliers, we calculated the following values.

$$D_{Q1} = max\{E_{med} - E_{Q1}, |.05 * E_{med}|\}$$

$$D_{Q3} = max\{E_{Q3} - E_{med}, |.05 * E_{med}|\}$$

$$E_{med} \ = the\ median\ value\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q1} \ = the\ first\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q3} = the\ third\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

The outliers follow outside the following range.

$$\{E_{med} - c_j * D_{Q1}, E_{med} + c_j * D_{Q3}\}$$

$$c_j = parameter\ that\ controls\ the\ width\ of\ the\ acceptance\ interval$$

We set three C values for each GQ type. The C values determined the bounds for review, suppress, and impute flags. We conducted a manual review by plotting x_i and y_i values to set the bounds by GQ type. Note, this review is somewhat subjective, but imitates common practice for establishment surveys.

Matching to IPEDS

Matching to CMS Nursing Home Data

References

Hidiroglou, M.A., and Berthelot, J.-M. (1986). ”Statistical Editing and Imputation for Periodic Business Surveys”. Survey Methodology, 12, 73-83.

Table 9: GQAC Expected Count by Imputation Status

Commented [PJC(F33]: Will Tables 8 - 11 be adjusted when we determine the total set of unresolved cases, including the implauible cases with a response > 0?

Commented [JEZ(F34R33]: Updated, same comment from table 6 applies.

GQAC Expected Count	Resolved	Unresolved	Total
Populated	91,000	11,000	102,000
Not Populated	89,000	35,500	125,000
Total	180,000	46,500	227,000

Table 10: GQAC Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	113,000	18,000	131,000
Not Populated	67,000	29,000	96,000
Total	180,000	46,500	227,000

Table 11: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	83,000	16,500	99,500
Not Populated	97,000	30,000	127,000
Total	180,000	46,500	227,000

Table 12: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total

Disclosure Prohibited. Title 13 U.S. Code

Populated	151,000	33,000	183,000
Not Populated	29,500	14,000	43,000
Total	180,000	46,500	227,000

Disclosure Prohibited. Title 13 U.S. Code

Group Quarters Imputation Methodology

Commented [JEZ(F1]: New version. Somehow I have locked myself out of version 3.

Commented [JEZ(F2R1]: Version 3 2 combines comments from John Abowd and Pat.

Table of Contents

Table of Tables

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, especially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic. Large GQs are often the only addresses in their tabulation census block. Consequently, information suggesting that such a GQ has very low population, or zero, will be evident in the PL 94-171 redistricting data summary file.

Commented [JEZ(F3]: Tables based on 12/18/20 data.

Commented [JEZ(F4R3]: Ryan will have a new file available 12/20/20. Once I get it I can re-run HB and recreate the tables in the doc.

Commented [JEZ(F5R3]: Research will use data as of 12/13/20. GEO counts will be used in production.

Commented [JMA(F6]: I suggest adding a table with the sources of the data used to classify the MAFID as "occupied group quarters." Be clear about which 2020 Census operation so classified each one. Summarize here.

Commented [JEZ(F7R6]: Need to ask Ryan and Debbie.

A special telephone operation was conducted to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation that pass an initial quality review as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that were open on Census Day, but were vacant during the GQ Enumeration visit (which occurred between late July and mid-October 2020) require imputation.

In addition, we will impute a population size for GQs that do not meet our quality edits, as implemented in the DRF1 review process. These GQs have a reported population that is not plausible and likely erroneous. The criteria for implausibility differ by GQ type. They are based on historical information maintained in the MAF database and other sources available to the DSSD, POP and SEHSD reviewers. They will be consistently applied to all GQ reports. For example, a facility, which is a collection of GQs operated by a single reporting organization, might have reported the entire facility population in one GQ MAFID. We will employ the Hidiroglou-Berthelot (HB) editing process, commonly used in establishment surveys, to identify these GQs with an implausible reported population. Information about HB edits is included in the Appendix.

This document details a joint research effort by staff in DSSD and CES to determine a method for GQ Count Imputation. A specification will be written to detail the final method that is implemented in production. A results memo will document the results of the production imputation. The data in this document represent the GQ Universe as of December 13, 2020. This universe formed the basis of our research into possible imputation methods. The universe in production will change as a result of the NPC calling operation and subsequent review by staff in GEO. Final imputation results will be provided to the POP division for subject-matter-review.

Commented [TLK(F8]: A separate results memo should document the results and include the final universe and counts. The research universe should be similar to the final universe, but they don't have to be exactly the same.
I do think it would be nice to fix the universe after the NPC effort since over 10,000 GQs were worked and could have an impact on the final results.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that were enumerated in a status related to occupied, but (1) do not have a reported count, or (2) have an implausible reported count. This universe is made up of GQs with a status of Occupied; Open on Census Day, but Vacant During Visit[1]; and Refusals. We consider these GQs unresolved and will impute a count for them. Some of the occupied GQs with a reported population were treated as unresolved because their census day

[1] During GQ enumeration, the GQ was found to be vacant, but the contact at the GQ said the GQ was open on Census Day. This is different from the vacant GQs which were reported to be vacant on Census Day.

population was implausible. The goal of the GQ Count Imputation is to determine a population count for all 43,000 occupied GQs with no reported population as well as the 3,500 occupied GQs with an implausible population count. We do not expect any unresolved GQs after implementing the GQ Count Imputation.

Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. The first three rows represent the occupied GQ universe. The other statuses of Vacant, Delete, and Nonresidential are considered out-of-scope for GQ Count Imputation. The GQ Count Imputation will only impute a positive population count.

Table 1: GQ Universe as of December 13, 2020

GQ Status	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Occupied GQ	177,000	17,000	3,100	197,000
Open on Census Day, Vacant During Visit	1,700	19,500	200	21,500
Refusal GQ	900	6,700	200	7,800
Vacant GQ	30,500	0	0	30,500
Delete GQ	7,600	0	0	7,600
Nonresidential GQ	2,500	0	0	2,500
Total	220,000	43,000	3,500	267,000

Commented [JEZ(F9]: Do we want to do any tables about the GEO review? Or will we just re-create this table after we incorporate the 'N's as resolved?

Commented [TLK(F10R9]: I think this table should be updated once the review is complete. Thus the N will be included in the resolved column.

Commented [PJC(F11]: I agree. And somewhere we can describe briefly the calling operation, the data collected in it, and GEO's review.

Commented [JEZ(F12R11]: See my comment above.

Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 7 in the Appendix has a full list of the GQ type codes.

Table 2: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

GQ Type	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Correctional Facilities*	13,000	2,800	250	16,000
Juvenile Facilities	6,100	1,800	90	8,000
Nursing Facilities*	24,500	3,200	550	28,500
Hospitals	1,900	800	70	2,800
College Housing*	29,000	5,500	1,300	36,000
Military*	3,000	1,900	70	5,000
Shelters	24,500	8,200	150	33,000
Group Homes	62,000	9,100	700	72,000
Other	16,000	9,700	300	26,000
Total	180,000	43,000	3,500	227,000

*denotes GQ Type is included in NPC calling operation

In order to avoid duplicated persons in the GQ MAFIDs that will receive imputed population counts, responses found among records identified in the unduplication as persons properly residing in one of the unresolved GQ MAFIDs will be assigned to that MAFID and subtracted from the imputed population. This avoids double counting such people.

Commented [JMA(F13]: We should ensure that some version of this sentence is implemented. I think the first part automatic in unduplication from DRF1 to DRF2. Not sure about the subtraction part. It will be important to actively document our unduplication efforts here. The HB algorithm also addresses this, but it is not transparent to a lay reader.

Commented [JEZ(F14R13]: Need to ask Ryan about this.

Commented [TLK(F15R13]: We should check with Ryan. It is also possible that we would impute fewer people than reported and would need to remove people.
One option would be to suppress all person responses if the GQ is imputed. Then, impute all whole-person records.

Imputation Methods

Commented [PJC(F16]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Added: As our method has developed, and we're embracing the median imputation option when no data are available, it appears that this set will be empty.

Commented [JMA(F17R16]: Agree with Pat.

Commented [TLK(F18R16]: I tried to make it clear that we will get an imputed value for all unresolved GQs.

Variables

Table 3 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, Administrative Records, and nursing home data from the Centers for Medicare & Medicaid Services (CMS). We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 3: Auxiliary and Historical Data at the GQ-Level

Commented [JEZ(F19]: Add nursing home data from CMS. MEPS data?

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR
Number of All Beds	Total number of resident beds in the facility as reported by the provider.	CMS
Number of Occupied Beds	Total number of resident beds that are currently occupied as reported by the provider.	CMS

Commented [JEZ(F20]: Have we dropped some of these? I think from what Andy showed today he was just using the 4 counts in the poisson? Maybe I heard that wrong.

Commented [TLK(F21R20]: We can still show all of the data we are considering, even if we don't use them all.

An additional source available for college housing GQs is the Integrated Postsecondary Education Data System (IPEDS). These variables are available at the facility level but not for individual MAFIDs.

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

Commented [JEZ(F22]: We could include info in the appendix regarding matching.

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) "**capacity utilization**"—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Table 4: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Data Editing

The HB edits employed to detect implausible population counts can also be used to determine which resolved GQs contribute to the imputation base. While the most extreme outliers are flagged for imputation, for less extreme outliers, we will accept the reported values, but keep those GQs from contributing to the statistical estimation that produces the imputation. We compared reported population counts with four auxiliary counts to flag outliers

- GQAC Expected Count
- GQAX Max Number of People
- Current GQ Size
- Max Number of People

If the ratio between the reported population count and the auxiliary population count is determined to be an outlier, both the reported count and auxiliary count are removed from the estimated models. Note, the HB edit takes the GQ size into account, the ratios are transformed so that more importance is placed on a small deviation from the median ratio for a large GQ as opposed to a large deviation for a small GQ (Hidiroglou and Berthelot, 1986). Table 5 shows counts of GQs that were suppressed from our imputation models for our research. The same GQ could be flagged for suppression by more than one outlying ratio, therefore the total number of suppressed GQs is not equal to the sum of the flags for each ratio.

Disclosure Prohibited. Title 13 U.S. Code

Table 5: Counts of GQs suppressed from imputation models by GQ Type

GQ Type	Suppressed from Models					
	GQAC Expected Count	GQAC Max Number of People	Current GQ Size	Max Number of People	Total Suppressed	Total Resolved
Correctional Facilities	N<15	N<15	90	90	150	13,000
Juvenile Facilities	20	30	40	60	90	6,100
Nursing Facilities	20	N<15	20	40	80	24,500
Hospitals	20	N<15	20	30	50	1,900
College Housing	250	N<15	50	80	400	29,000
Military	N<15	20	20	N<15	50	3,000
Shelters	30	N<15	30	100	150	24,500
Group Homes	60	N<15	150	30	200	62,000
Other	30	20	30	70	150	16,000
Total	450	150	450	500	1,300	180,000

In addition to outliers flagged for imputation or to be eliminated from the imputation base for our models, a third set of less extreme outliers will be identified with a flag for review. These flags may help to prioritize subject-matter review of final GQ counts.

Candidate Methods

For GQ Count Imputation we evaluated the following methods:

1. Ratio Imputation
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Percentile Imputation

Commented [ADK(F23]: Need to look at paradata as covariates as well on the models

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ population count, as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we can use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error relative to the Census Day target of April 1 than other methods.

Table 5 shows that 11,000 of the unresolved GQs included in our research could be resolved by converting the GQAC expected count to the GQ pop count using the ratio adjustment.

Table 6: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	91,000	11,000	192,000
Not Populated	89,000	35,500	125,000
Total	180,000	46,500	227,000

Commented [JEZ(F24]: Updated with the latest.

Commented [JEZ(F25R24]: We should think about what to do in production when we have a flag on the GP/Exp count ratio - we don't know which value is 'wrong'. We are trying to impute GP, so we could use expected count, but we don't want to do that if the expected count is what is off in the ratio. For the truth deck and for the donor pool, I think it's fine to throw out both but for production we need to figure out when we should accept GP or expected count (same applies to the other vars). For now, no rule has been applied to this table (i.e. ID could be unresolved because of an I flag on the GP/expected count ratio and still have expected count populated in this table).

For each detailed GQ type (see Table 8) within each state , we used the resolved cases with a GQAC expected count to calculate the ratio of the reported GQ Census Day count to the GQAC expected count.

We then used this ratio to convert the GQAC expected count of the unresolved GQs into a Census Day imputed count. For example, for an unresolved College GQ with a GQAC Expected Count, the following equation would be applied:

$$Imputed\ Population\ Count = \ GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We constructed ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. Ratios for the four variables by GQ Type are shown in Table 6. The ratios presented here were not used directly – we used the more detailed GQ type and state to calculate the ratios in the imputation. Table 9 through Table 11 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 7: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Reported Count to GQAC Expected Count	Ratio of Reported Count to GQAC Max Number of People	Ratio of Reported Count to Current GQ Size	Ratio of Reported Count to Max Number of People
Correctional Facilities	0.7389	0 6613	0.8960	0.7328
Juvenile Facilities	0.7363	0.5870	0.8713	0.6636
Nursing Facilities	0.8710	0.7482	0.7925	0.5916
Hospitals	0.7925	0 6928	0.9360	0.7463
College Housing	0.9056	0.8069	0.8986	0.6881
Military	0.7540	0.6822	0.9249	0.8118
Shelters	0.6353	0.6115	0.8652	0.5490
Group Homes	0.8816	0.7792	0.6502	0.6703
Other	0.8453	0.6137	0 9216	0.7851
All GQs	0.8480	0.7342	0.8960	0.7320

Commented [JEZ(F26]: Removed all IDs for which ANY flags are 'S' or 'I'. This lines up with what Andy is using in the truth deck. In production, we probably will want to only exclude for certain ratios (i e.if GP/Exp count looks okay, keep in that ratio but GP/Max count is flagged, exclude from that ratio)

Commented [JEZ(F27]: Do we still want the Greek break-out? Need to ask Andy if he is using it.

Adjusted Residual from Facility-level Total for College Housing

The second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for colleges and universities (GQTYPCUR=501).

First, we adjusted the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population, GQAC Max Number of People, and Greek Housing variables.

After adjusting the college-level total room capacity to account for reference year and capacity utilization, we calculated the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count$$
$$- \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Commented [TLK(F28]: From John Abowd: We should call out that the Reported GQ Population Count contains the persons who were properly placed in the GQ during unduplication.

Once we calculated the college-level residual, we then allocated the population counts among the GQs in the college without GQAC Expected Count.

Modeling

The third approach would be to impute the GQ population counts from a Poisson regression model. The dependent variable would be the natural logarithm of the ratio of reported GQ population count to GQ Current Max Size. Independent variables are

- GQ Type
- GQAC Expected Count
- GQAC Max Number of People
- Current GQ Size

See Table 3 for a description of the covariates. It is important to note that GQ type is a fixed-effect covariate in the model. Each model will contain the same set of covariates, except for the college model, which will also include an indicator for Greek Housing.

Commented [TLK(F29]: Check with Andy.

Percentile Imputation

If sufficient auxiliary data are not available, we will impute the population count with the median population count of the resolved GQs within detailed GQ type and state. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and state. Then, we will calculate the median or other percentile of the GQ population count and impute the unresolved GQs with the median or other percentile of the GQ population count in the cell. We will determine the percentile to use based on the value that minimizes the imputation bias in our evaluation.

Evaluation of Imputed Values

We evaluated the imputation methods using 10-fold cross validation. First, we removed the unresolved GQs from the universe since we don't have a reported GQ population count for them. Next, we removed GQs with a count that was implausible based on our edits. We also removed any GQs that had any of the four flags set to suppress the reported population count from the imputation models. This ensured that extreme outliers would not influence our evaluation. We then divided the remaining resolved occupied GQs into 10 approximately equal sized groups.

To create these groups, we sampled at the tract- and unit-levels. To ensure that we had certain tracts without missing GQs the list of tracts containing GQs was split in half. The first half was sampled at the tract-level. We employed systematic sampling, ordering these tracts by state, county, and then tract. The second half of the list was sampled at the unit-level. The units within these tracts were ordered by tract, GQ type, and GQ(MAFID?) and again, we employed systematic sampling to allocate each GQ to a group. Each of the ten groups had approximately 18 000 GQs.

Commented [JEZ(F30]: I think this is what Andy is saying in his TD memo. Need to check.

We built and fit our models on nine of the groups and then imputed responses for the remaining group, repeating this process 10 times, so that an out-of-sample forecast error is available for each observation in the imputation base used for estimation. We evaluated all four methods, imputing as many units in the "unresolved" group as possible.

Then, we calculated the difference between the reported GQ population count and the 10-fold predicted value for each of the four imputation methods. This generated the out-of-sample forecast errors. The average of these differences is an estimate of the 10-fold cross validation prediction error. Since the distribution of GQ population sizes is skewed, we will also calculate the median and interquartile range of the differences.

$$Forecast\ Error_{Mi} = \frac{\sum_{k}^{10}(Reported\ Population\ Count_i - Predicted\ Value_{Mik})}{10}$$

$$Prediction\ Error_M = \frac{\sum_{i}^{N} Forecast\ Error_{Mi}}{N}$$

$$where\ M = method, k = group, i = observation, N = total\ number\ of\ observations$$

Some methods may only work under certain conditions. For example, the IPEDS residual method will only work for colleges. The Poisson regression will only work when all of the necessary covariates are filled. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Research Results

We first examined the cases where all four auxiliary counts were available, in order to compare the Poisson method with the other candidate methods. Note, 5,100 of the unresolved cases as of December 13, 2020 have all four auxiliary counts populated. The GQs with implausible counts are not included in this total since at least one of the ratios comparing the reported population count and the auxiliary count is an outlier. This could mean the reported population count is erroneous or that the auxiliary count is erroneous. For 900 of these 5,100 GQs, all auxiliary counts are equal.

For the resolved IDs, 65,000 have all auxiliary counts populated. Of these, 13,500 have equal values for all four auxiliary counts. Of these, 7,000 have reported counts equal to the auxiliary counts, 4,100 have reported counts less than auxiliary counts, and 2,700 have reported counts greater than the auxiliary counts.

Commented [TLK(F31]: We should try to compute standard errors of the 10-fold cross validation prediction error as described in Rodriquez. However, I think it would also be useful to look at the range and interquartile range of the differences.

Commented [TLK(F32]: I think this section is correct.
I updated it based on John Abowd's comment:
This is not the correct procedure for k-fold cross validation. If N is the total number of observations, then for each i in N, you have exactly one k-fold forecast error for each method. Then number of groups (k=10) is no longer relevant. You compute the statistics for each of the four methods from the N out-of-sample forecast errors. The method described in the text is only correct if the aggregate error measure is linear in the out-of-sample forecasts. Details of for other error measures can be found in the Rodriquez et al (2010) article I put in the same directory as these notes.

Commented [JEZ(F33]: Am I understanding this correctly?

Commented [JEZ(F34]: Need to add some table shells. Even if they're not filled in on Wednesday we can get some direction.

Commented [JEZ(F35]: I don't know that we need to include all of this, I was just looking at these today.

Table 8: Prediction Error for each candidate method by GQ Type.

GQ Type	Ratio Adjusted By 2020 GQAC Expected Count	Ratio Adjusted By 2020 GQAC Max Count	Ratio Adjusted By 2020 Current GQ Size	Ratio Adjusted By 2020 Max Number of People	Poisson Model	Precentile
Correctional Facilities	12.34 (5.67)					
Juvenile Facilities	12.34 (5.67)					
Nursing Facilities						
Hospitals						
College Housing						
Military						
Shelters						
Group Homes						
Other						
All GQs						

Standard errors are in parentheses.

Commented [JEZ(F36]: Make separate table for CES methods after we determine order for the first 6 methods. Need to write this up in evaluation.

Recommendation

Appendix

Table 9: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities

Disclosure Prohibited. Title 13 U.S. Code

CODE	VALUE
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Hidiroglou-Berthelot (HB) Edits

The Hidiroglou-Berthelot (HB) edit detects outliers based on the ratio of two variables. In calculating the HB statistic, the ratio is transformed once to ensure that outliers are identified at both tails of the HB statistic's distribution, then transformed again to account for the size of the reporting unit. This results in identifying the records whose data exhibit the most unusual differences between the numerator and denominator variables as well as those that have more impact on the totals. These data are identified as requiring analyst review, suppression from the imputation base, or imputation.

For our purposes in this project, the HB statistic was calculated as follows.

First, we calculated the ratio between the reported population count and the auxiliary population count for each GQ with positive counts for both values.

$$R_i = {}^{x_i}/_{y_i}$$

$$x_i \; = Reported\ Population\ Count$$

$$y_i = Auxiliary\ Population\ Count$$

We then transformed the ratios in order to detect outliers at both tails of the distribution. We calculated median ratios within each GQ type.

Disclosure Prohibited. Title 13 U.S. Code

$$S_i = \begin{cases} 1 - \dfrac{R_{med}}{R_i} & 0 < R_i < R_{med} \\ \dfrac{R_i}{R_{med}} - 1 & R_i > R_{med} \end{cases}$$

$$R_{med} = median\ R_i$$

We then scaled the transformed ratios by GQ size to calculate the HB statistic.

$$E_i = S_i * \sqrt{\{\max(x_i, y_i)\}}$$

To detect outliers, we calculated the following values.

$$D_{Q1} = max\{E_{med} - E_{Q1}, |.05 * E_{med}|\}$$

$$D_{Q3} = max\{E_{Q3} - E_{med}, |.05 * E_{med}|\}$$

$$E_{med} = the\ median\ value\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q1} = the\ first\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q3} = the\ third\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

The outliers fall outside the following range.

$$\{E_{med} - c_j * D_{Q1}, E_{med} + c_j * D_{Q3}\}$$

$$c_j = parameter\ that\ controls\ the\ width\ of\ the\ acceptance\ interval$$

We set three C values for each GQ type. The C values determined the bounds for review, suppress, and impute flags. We conducted a manual review by plotting x_i and y_i values to set the bounds by GQ type. Note, this review is somewhat subjective, but imitates common practice for establishment surveys.

Matching to IPEDS

Matching to CMS Nursing Home Data

References

Hidiroglou, M.A., and Berthelot, J.-M. (1986). "Statistical Editing and Imputation for Periodic Business Surveys". Survey Methodology, 12, 73-83.

Table 10: GQAC Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	91,000	11,000	192,000
Not Populated	89,000	35,500	125,000
Total	180,000	46,500	227,000

Commented [PJC(F37]: Will Tables 8 - 11 be adjusted when we determine the total set of unresolved cases, including the implauible cases with a response > 0?

Commented [JEZ(F38R37]: Updated, same comment from table 6 applies.

Table 11: GQAC Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	113,000	18,000	131,000
Not Populated	67,000	29,000	96,000
Total	180,000	46,500	227,000

Table 12: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	83,000	16,500	99,500
Not Populated	97,000	30,000	127,000
Total	180,000	46,500	227,000

Table 13: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	151,000	33,000	183,000
Not Populated	29,500	14,000	43,000
Total	180,000	46,500	227,000

Disclosure Prohibited. Title 13 U.S. Code

Group Quarters Imputation Methodology

Commented [JEZ(F1]: New version. Somehow I have locked myself out of version 3.

Commented [JEZ(F2R1]: Version 3 2 combines comments from John Abowd and Pat.

Table of Contents

Table of Tables

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, especially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic. Large GQs are often the only addresses in their tabulation census block. Consequently, information suggesting that such a GQ has very low population, or zero, will be evident in the PL 94-171 redistricting data summary file.

Commented [JEZ(F3]: Tables based on 12/18/20 data.

Commented [JEZ(F4R3]: Ryan will have a new file available 12/20/20. Once I get it I can re-run HB and recreate the tables in the doc.

Commented [JEZ(F5R3]: Research will use data as of 12/13/20. GEO counts will be used in production.

Commented [JMA(F6]: I suggest adding a table with the sources of the data used to classify the MAFID as "occupied group quarters." Be clear about which 2020 Census operation so classified each one. Summarize here.

Commented [JEZ(F7R6]: Need to ask Ryan and Debbie.

A special telephone operation was conducted to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation that pass an initial quality review as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that were open on Census Day, but were vacant during the GQ Enumeration visit (which occurred between late July and mid-October 2020) require imputation.

In addition, we will impute a population size for GQs that do not meet our quality edits, as implemented in the DRF1 review process. These GQs have a reported population that is not plausible and likely erroneous. The criteria for implausibility differ by GQ type. They are based on historical information maintained in the MAF database and other sources available to the DSSD, POP and SEHSD reviewers. They will be consistently applied to all GQ reports. For example, a facility, which is a collection of GQs operated by a single reporting organization, might have reported the entire facility population in one GQ MAFID. We will employ the Hidiroglou-Berthelot (HB) editing process, commonly used in establishment surveys, to identify these GQs with an implausible reported population. Information about HB edits is included in the Appendix.

This document details a joint research effort by staff in DSSD and CES to determine a method for GQ Count Imputation. A specification will be written to detail the final method that is implemented in production. A results memo will document the results of the production imputation. The data in this document represent the GQ Universe as of December 13, 2020. This universe formed the basis of our research into possible imputation methods. The universe in production will change as a result of the NPC calling operation and subsequent review by staff in GEO. Final imputation results will be provided to the POP division for subject-matter-review.

Commented [TLK(F8]: A separate results memo should document the results and include the final universe and counts. The research universe should be similar to the final universe, but they don't have to be exactly the same.
I do think it would be nice to fix the universe after the NPC effort since over 10,000 GQs were worked and could have an impact on the final results.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that were enumerated in a status related to occupied, but (1) do not have a reported count, or (2) have an implausible reported count. This universe is made up of GQs with a status of Occupied; Open on Census Day, but Vacant During Visit[1]; and Refusals. We consider these GQs unresolved and will impute a count for them. Some of the occupied GQs with a reported population were treated as unresolved because their census day

[1] During GQ enumeration, the GQ was found to be vacant, but the contact at the GQ said the GQ was open on Census Day. This is different from the vacant GQs which were reported to be vacant on Census Day.

population was implausible. The goal of the GQ Count Imputation is to determine a population count for all 43,000 occupied GQs with no reported population as well as the 2,200 occupied GQs with an implausible population count. We do not expect any unresolved GQs after implementing the GQ Count Imputation.

Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. The first three rows represent the occupied GQ universe. The other statuses of Vacant, Delete, and Nonresidential are considered out-of-scope for GQ Count Imputation. The GQ Count Imputation will only impute a positive population count.

Table 1: GQ Universe as of December 13, 2020

GQ Status	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Occupied GQ	179,000	17,000	1,900	197,000
Open on Census Day, Vacant During Visit	1,700	19,500	200	21,500
Refusal GQ	900	6,700	150	7,800
Vacant GQ	30,500	0	0	30,500
Delete GQ	7,600	0	0	7,600
Nonresidential GQ	2,500	0	0	2,500
Total	220,000	43,000	2,200	267,000

Commented [JEZ(F9]: Do we want to do any tables about the GEO review? Or will we just re-create this table after we incorporate the 'N's as resolved?

Commented [TLK(F10R9]: I think this table should be updated once the review is complete. Thus the N will be included in the resolved column.

Commented [PJC(F11]: I agree. And somewhere we can describe briefly the calling operation, the data collected in it, and GEO's review.

Commented [JEZ(F12R11]: See my comment above.

Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 7 in the Appendix has a full list of the GQ type codes.

Table 2: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

GQ Type	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Correctional Facilities*	13,000	2,800	150	16,000
Juvenile Facilities	6,100	1,800	60	8,000
Nursing Facilities*	25,000	3,200	450	28,500
Hospitals	1,900	800	60	2,800
College Housing*	29,500	5,500	650	36,000
Military*	3,100	1,900	40	5,000
Shelters	24,500	8,200	100	33,000
Group Homes	62,000	9,100	500	72,000
Other	16,000	9,700	200	26,000
Total	181,000	43,000	2,200	227,000

*denotes GQ Type is included in NPC calling operation

In order to avoid duplicated persons in the GQ MAFIDs that will receive imputed population counts, responses found among records identified in the unduplication as persons properly residing in one of the unresolved GQ MAFIDs will be assigned to that MAFID and subtracted from the imputed population. This avoids double counting such people.

Commented [JMA(F13]: We should ensure that some version of this sentence is implemented. I think the first part automatic in unduplication from DRF1 to DRF2. Not sure about the subtraction part. It will be important to actively document our unduplication efforts here. The HB algorithm also addresses this, but it is not transparent to a lay reader.

Commented [JEZ(F14R13]: Need to ask Ryan about this.

Commented [TLK(F15R13]: We should check with Ryan. It is also possible that we would impute fewer people than reported and would need to remove people.
One option would be to suppress all person responses if the GQ is imputed. Then, impute all whole-person records.

Imputation Methods

Commented [PJC(F16]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Added: As our method has developed, and we're embracing the median imputation option when no data are available, it appears that this set will be empty.

Commented [JMA(F17R16]: Agree with Pat.

Commented [TLK(F18R16]: I tried to make it clear that we will get an imputed value for all unresolved GQs.

Variables

Table 3 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, Administrative Records, and nursing home data from the Centers for Medicare & Medicaid Services (CMS). We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 3: Auxiliary and Historical Data at the GQ-Level

Commented [JEZ(F19]: Add nursing home data from CMS. MEPS data?

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR
Number of All Beds	Total number of resident beds in the facility as reported by the provider.	CMS
Number of Occupied Beds	Total number of resident beds that are currently occupied as reported by the provider.	CMS

Commented [JEZ(F20]: Have we dropped some of these? I think from what Andy showed today he was just using the 4 counts in the poisson? Maybe I heard that wrong.

Commented [TLK(F21R20]: We can still show all of the data we are considering, even if we don't use them all.

An additional source available for college housing GQs is the Integrated Postsecondary Education Data System (IPEDS). These variables are available at the facility level but not for individual MAFIDs.

We have the college-level total room capacity (number of persons that could live in the GQ) from the IPEDS for the 2019-2020 academic year. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least two reasons:

Commented [JEZ(F22]: We could include info in the appendix regarding matching.

(~~2~~1) **"capacity utilization"**—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

Disclosure Prohibited. Title 13 U.S. Code

(~~3~~2) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Table 4: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Data Editing

The HB edits employed to detect implausible population counts can also be used to determine which resolved GQs contribute to the imputation base. While the most extreme outliers are flagged for imputation, for less extreme outliers, we will accept the reported values, but keep those GQs from contributing to the statistical estimation that produces the imputation. We compared reported population counts with four auxiliary counts to flag outliers

- GQAC Expected Count
- GQAX Max Number of People
- Current GQ Size
- Max Number of People

If the ratio between the reported population count and the auxiliary population count is determined to be an outlier, both the reported count and auxiliary count are removed from the estimated models. Note, the HB edit takes the GQ size into account, the ratios are transformed so that more importance is placed on a small deviation from the median ratio for a large GQ as opposed to a large deviation for a small GQ (Hidiroglou and Berthelot, 1986). Table 5 shows counts of GQs that were suppressed from our imputation models for our research. The same GQ could be flagged for suppression by more than one outlying ratio, therefore the total number of suppressed GQs is not equal to the sum of the flags for each ratio.

Disclosure Prohibited. Title 13 U.S. Code

Table 5: Counts of GQs suppressed from imputation models by GQ Type

GQ Type	Suppressed from Models					
	GQAC Expected Count	GQAC Max Number of People	Current GQ Size	Max Number of People	Total Suppressed	Total Resolved
Correctional Facilities	20	50	150	150	250	13,000
Juvenile Facilities	40	60	40	80	100	6,100
Nursing Facilities	80	40	90	80	200	25,000
Hospitals	20	20	20	40	60	1,900
College Housing	700	350	200	600	1,000	29,500
Military	20	40	30	30	80	3,100
Shelters	50	20	40	150	200	24,500
Group Homes	200	100	150	200	450	62,000
Other	70	90	50	100	250	16,000
Total	1,200	800	800	1,400	2,600	181,000

In addition to outliers flagged for imputation or to be eliminated from the imputation base for our models, a third set of less extreme outliers will be identified with a flag for review. These flags may help to prioritize subject-matter review of final GQ counts.

Candidate Methods

For GQ Count Imputation we evaluated the following methods:

1. Ratio Imputation
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Percentile Imputation

Commented [ADK(F23]: Need to look at paradata as covariates as well on the models

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ population count, as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we can use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error relative to the Census Day target of April 1 than other methods.

Table 5 shows that 11,000 of the unresolved GQs included in our research could be resolved by converting the GQAC expected count to the GQ population count using the ratio adjustment.

Table 6: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	92,000	10,000	102,000
Not Populated	89,000	35,500	125,000
Total	181,000	45,500	227,000

For each detailed GQ type (see Table 8) within each state, we used the resolved cases with a GQAC expected count to calculate the ratio of the reported GQ Census Day count to the GQAC expected count.

We then used this ratio to convert the GQAC expected count of the unresolved GQs into a Census Day imputed count. For example, for an unresolved College GQ with a GQAC Expected Count, the following equation would be applied:

$$Imputed\ Population\ Count = \ GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We constructed ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. Ratios for the four variables by GQ Type are shown in Table 6. The ratios presented here were not used directly – we used the more detailed GQ type and state to calculate the ratios in the imputation. Table 9 through Table 11 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 7: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Reported Count to GQAC Expected Count	Ratio of Reported Count to GQAC Max Number of People	Ratio of Reported Count to Current GQ Size	Ratio of Reported Count to Max Number of People
Correctional Facilities	0.7387	0 6551	0.8692	0.6661
Juvenile Facilities	0.7388	0.5862	0.7962	0.5885
Nursing Facilities	0.8705	0.7442	0.9365	0.7432
Hospitals	0.8017	0 6970	0.9035	0.6942
College Housing	0.9051	0.7890	0.9235	0.7999
Military	0.7552	0.6862	0.8672	0.5513
Shelters	0.6324	0.6097	0.6521	0.6678
Group Homes	0.8796	0.7676	0.9197	0.7784
Other	0.8499	0.6044	0.7906	0.6026
All GQs	0.8469	0.7249	0.8953	0.7285

Adjusted Residual from Facility-level Total for College Housing

The second imputation method under consideration is ~~the~~ a hybrid approach where we use GQ-level auxiliary variables for imputation when they are available and allocate an Adjusted Residual from Facility-level Totals for College Housing when the GQ-level variables are not available~~.~~ This method can only be used for colleges and universities (GQTYPCUR=501).

First, ~~we adjusted the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population,~~ we impute GQ-level population counts using ~~GQAC Max Number of People~~the GQ-level auxiliary variable and a ratio method similar to one described in the previos section. One potentially important difference is that we calculate college-specific ratios when possible.~~ and Greek Housing variables~~ If none of the GQ-level auxiliary population variables is available we use a college- or state-specific median for GQs that are Greek houses.

After adjusting the college-level total room capacity to account for reference year and capacity utilization, we calculated the following college-level residual for each college C:

$$\begin{aligned} Residual_C = {} & \cancel{Adjusted\ } IPEDS\ Room\ Capacity_C \\ & - \sum_{gq \in S1 \cancel{C}} \cancel{Reported\ GQ\ Pop\ Count} Max\ Number\ of\ People_{gq} \\ & - \sum_{gq \in S2} Reported\ Pop\ Count_{gq} \\ & - \sum_{\cancel{C \notin} gq \in S3} \cancel{GQAC\ Expected\ Count} Imputed\ Pop\ Count_{gq} \end{aligned}$$

where the first summation is over all GQs at college C with a good number for Maximum Number of People, ~~person count~~, ~~and~~ the second summation is over all GQs at college C *without* a good ~~person count~~Maximum Number of Persons, but with ~~positive~~ a good *positive* reported population count ~~GQAC Expected Count~~ and the third summation is over GQs which do not fall in the first two groups and for which we have already imputed a population count.

Formatted: Font: Italic

Commented [TLK(F24]: From John Abowd: We should call out that the Reported GQ Population Count contains the persons who were properly placed in the GQ during unduplication.

Once we have calculated the college-level residual, we then allocate~~d~~ the residual ~~population counts~~ among the GQs in the college~~s~~ without ~~GQAC Expected Count~~ a good reported population count or an already-imputed population count. As a fallback, if the college-level residual is negative or if the allocated residual is larger than the largest repoted population count at this college, then we impute using the median reported population among non-Greek GQs at this college or at colleges in the same state.

Modeling

The third approach would be to impute the GQ population counts from a Poisson regression model. The dependent variable would be the natural logarithm of the ratio of reported GQ population count to GQ Current Max Size. Independent variables are

- GQ Type
- GQAC Expected Count
- GQAC Max Number of People
- Current GQ Size

See Table 3 for a description of the covariates. It is important to note that GQ type is a fixed-effect covariate in the model. Each model will contain the same set of covariates.~~, except for the college model, which will also include an indicator for Greek Housing.~~

Commented [TLK(F25]: Check with Andy.

Commented [JEZ(F26R25]: He's not using it.

Percentile Imputation

If sufficient auxiliary data are not available, we will impute the population count with the median population count of the resolved GQs within detailed GQ type and state. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and state. Then, we will calculate the median or other percentile of the GQ population count and impute the unresolved GQs with the median or other percentile of the GQ population count in the cell. We will determine the percentile to use based on the value that minimizes the imputation bias in our evaluation.

Evaluation of Imputed Values

We evaluated the imputation methods using 10-fold cross validation. First, we removed the unresolved GQs from the universe since we don't have a reported GQ population count for them. Next, we removed GQs with a count that was implausible based on our edits. We also removed any GQs that had any of the four flags set to suppress the reported population count from the imputation models. This ensured that extreme outliers would not influence our evaluation. We then divided the remaining resolved occupied GQs into 10 approximately equal sized groups.

To create these groups, we sampled at the tract- and unit-levels. To ensure that we had certain tracts without missing GQs, the list of tracts containing GQs was randomly sorted and then split in half. The first half was sampled at the tract-level. We employed systematic sampling, ordering these tracts by state, county, and then tract. The second half was sampled at the unit-level. The units within these tracts were ordered by tract, GQ type, and GQ(MAFID?) and again, we employed systematic sampling to allocate each GQ to a group. Each of the ten groups had approximately 18,000 GQs.

Commented [JEZ(F27]: I think this is what Andy is saying in his TD memo. Need to check.

We built and fit our models on nine of the groups and then imputed responses for the remaining group, repeating this process 10 times, so that an out-of-sample forecast error is available for each observation in the imputation base used for estimation. We evaluated all four methods, imputing as many units in the "unresolved" group as possible.

Then, we calculated the difference between the reported GQ population count and the 10-fold predicted value for each of the four imputation methods. This generated the out-of-sample forecast errors. The average of these differences is an estimate of the 10-fold cross validation prediction error. Since the distribution of GQ population size is skewed, we will also calculate the median and interquartile range of the differences. The estimated prediction error is defined as.

$$Prediction\ Error_M = \frac{1}{N}\sum_{k=1}^{10}\sum_{i=1}^{N_k}(Reported\ Population\ Count_i - Predicted\ Value_{Mik})$$

$$where\ M = method, k = group, i = observation, N = total\ number\ of\ observations$$

Commented [TLK(F28]: We should try to compute standard errors of the 10-fold cross validation prediction error as described in Rodriquez. However, I think it would also be useful to look at the range and interquartile range of the differences.

Commented [TLK(F29]: I think this section is correct. I updated it based on John Abowd's comment:
This is not the correct procedure for k-fold cross validation. If N is the total number of observations, then for each i in N, you have exactly one k-fold forecast error for each method. Then number of groups (k=10) is no longer relevant. You compute the statistics for each of the four methods from the N out-of-sample forecast errors. The method described in the text is only correct if the aggregate error measure is linear in the out-of-sample forecasts. Details of for other error measures can be found in the Rodriquez et al (2010) article I put in the same directory as these notes.

Some methods may only work under certain conditions. For example, the IPEDS residual method will only work for colleges. The Poisson regression will only work when all of the necessary covariates are filled. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Research Results

We first examined the cases where all four auxiliary counts were available, in order to compare the Poisson method with the other candidate methods. Note, 5,100 of the unresolved cases as of December 13, 2020 have all four auxiliary counts populated. The GQs with implausible counts are not included in this total since at least one of the ratios comparing the reported population count and the auxiliary count is an outlier. This could mean the reported population count is erroneous or that the auxiliary count is erroneous. For 900 of these 5,100 GQs, all auxiliary counts are equal.

Commented [JEZ(F30]: Need to add some table shells. Even if they're not filled in on Wednesday we can get some direction.

Commented [TLK(F31R30]: It might be better to plot the results in a graphic.

For the resolved IDs, 65,000 have all auxiliary counts populated. Of these, 14,000 have equal values for all four auxiliary counts. Of these, 7,000 have reported counts equal to the auxiliary counts, 4,100 have reported counts less than auxiliary counts, and 2,700 have reported counts greater than the auxiliary counts.

Commented [JEZ(F32]: I don't know that we need to include all of this, I was just looking at these today.

Recommendation

Appendix

Table 8: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults

CODE	VALUE
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Hidiroglou-Berthelot (HB) Edits

The Hidiroglou-Berthelot (HB) edit detects outliers based on the ratio of two variables. In calculating the HB statistic, the ratio is transformed once to ensure that outliers are identified at both tails of the HB statistic's distribution, then transformed again to account for the size of the reporting unit. This results in identifying the records whose data exhibit the most unusual differences between the numerator and denominator variables as well as those that have more impact on the totals. These data are identified as requiring analyst review, suppression from the imputation base, or imputation.

For our purposes in this project, the HB statistic was calculated as follows.

First, we calculated the ratio between the reported population count and the auxiliary population count for each GQ with positive counts for both values.

$$R_i = {}^{x_i}/_{y_i}$$

$$x_i \; = Reported\ Population\ Count$$

$$y_i = Auxiliary\ Population\ Count$$

We then transformed the ratios in order to detect outliers at both tails of the distribution. We calculated median ratios within each GQ type.

$$S_i = \begin{cases} 1 - \dfrac{R_{med}}{R_i} & 0 < \; R_i < \; R_{med} \\ \dfrac{R_i}{R_{med}} - 1 & R_i > R_{med} \end{cases}$$

$$R_{med} \; = median\ R_i$$

We then scaled the transformed ratios by GQ size to calculate the HB statistic.

$$E_i \; = \; S_i * \sqrt{\{\max{(x_i, y_i)}\}}$$

To detect outliers, we calculated the following values.

$$D_{Q1} = max\{E_{med} - E_{Q1}, |.05 * E_{med}|\}$$

$$D_{Q3} = max\{E_{Q3} - E_{med}, |.05 * E_{med}|\}$$

$$E_{med} \; = the\ median\ value\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q1} = the\ first\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q3} = the\ third\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

The outliers fall outside the following range.

$$\{E_{med} - c_j * D_{Q1}, E_{med} + c_j * D_{Q3}\}$$

$$c_j = parameter\ that\ controls\ the\ width\ of\ the\ acceptance\ interval$$

We set three C values for each GQ type. The C values determined the bounds for review, suppress, and impute flags. We conducted a manual review by plotting x_i and y_i values to set the bounds by GQ type. Note, this review is somewhat subjective, but imitates common practice for establishment surveys.

Matching to IPEDS

Matching to CMS Nursing Home Data

References

Hidiroglou, M.A., and Berthelot, J.-M. (1986). ”Statistical Editing and Imputation for Periodic Business Surveys”. Survey Methodology, 12, 73-83.

Table 9: GQAC Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	92,000	10,000	102,000
Not Populated	89,000	35,500	125,000
Total	181,000	45,500	227,000

Table 10: GQAC Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	114,000	16,500	131,000
Not Populated	67,500	28,500	96,000
Total	181,000	45,500	227,000

Table 11: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	83,500	16,000	99,500
Not Populated	97,500	29,500	127,000
Total	181,000	45,500	227,000

Table 12: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total

Disclosure Prohibited. Title 13 U.S. Code

Populated	152,000	31,500	183,000
Not Populated	29,500	14,000	43,000
Total	181,000	45,500	227,000

This Document Contains Title-13 Data

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A special telephone operation was conducted to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation that pass an initial quality review as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that open on Census Day, but vacant during the GQ Enumeration visit (which started in July 2020) require imputation.

In addition, we will impute a pop size for GQs that do not meet our quality edits. These GQs have a reported population that is not plausible and likely a error. For example, a facility might have reported their facilty population in one GQ. We employed the Hidiroglou-Berthelot (HB) editing process, commonly used in establishment surveys, to identify these GQs with an implausible reported population. Information about HB edits is included in the Appendix.

Commented [JEZ(F1]: Tables based on 12/18/20 data.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have an unplausible reported count. This universe is made up of GQs with a status of Occupied; Open on Census Day, but Vacant During Visit[1]; and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with implausible population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe. The other statuses of Vacant, Delete, and Nonresidential are considered out-of-scope for GQ Count Imputation. The GQ Count Imputaiton will only impute a positive population.

Table 1: GQ Universe

GQ Status	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Occupied GQ	177,000	17,000	3,100	197,000
Open on Census Day, Vacant During Visit	1,700	19,500	200	21,500
Refusal GQ	900	6,700	200	7,800
Vacant GQ	30,500	0	0	30,500
Delete GQ	7,600	0	0	7,600
Nonresidential GQ	2,500	0	0	2,500
Total	220,000	43,000	3,500	267,000

Commented [JEZ(F2]: Do we want to do any tables about the GEO review? Or will we just re-create this table after we incorporate the 'N's as resolved?

Commented [TLK(F3R2]: I think this table should be updated once the review is complete. Thus the N will be included in the resolved column.

[1] During GQ enumeration, the GQ was found to be vacant, but the contact at the GQ said the GQ was open on Census Day. This is different from the vacant GQs which were reported to be vacant on Census Day.

Some of the occupied GQs with a reported population were treated as unresolved because their census day population was implausible. The goal of the GQ Count Imputation is to determine a population count for all 43,000 occupied GQs with no reported population as well as the 3,500 occupied GQs with an implausible population count.

Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 7 in the Appendix has a full list of the GQ type codes.

Table 2: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs

GQ Type	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Correctional Facilities*	13,000	2,800	250	16,000
Juvenile Facilities	6,100	1,800	90	8,000
Nursing Facilities*	24,500	3,200	550	28,500
Hospitals	1,900	800	70	2,800
College Housing*	29,000	5,500	1,300	36,000
Military*	3,000	1,900	70	5,000
Shelters	24,500	8,170	166	32,769
Group Homes	62,000	9,100	700	72,000
Other	16,000	9,700	300	26,000
Total	180,000	43,000	3,500	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Commented [PJC(F4]: Do we still need some material at the end of the previous sections that indicates for which cases we will not impute? I'm thinking of cases for which we have no good auxiliary data on which to base the imputation. Will there be such cases?

Variables

Table 3 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, Administrative Records, and nursing home data from the Centers for Medicare & Medicaid Services (CMS). We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 3: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR
Number of All Beds	Total number of resident beds in the facility as reported by the provider.	CMS
Number of Occupied Beds	Total number of resident beds that are currently occupied as reported by the provider.	CMS

Commented [JEZ(F5]: From chat in EGG meeting: use 5-year ACS estimates?

Commented [JEZ(F6R5]: From James Christy: FWIW - when we visit a GQ for ACS, we ask for the total population of that GQ, then use that to sub-sample for selecting cases for interview. Reference date is when we visit the GQ. (Unlike Decennial which has a fixed reference date). I don't think that total pop count is part of what's published - but could be wrong.

Commented [TLK(F7R5]: Stuart Irby confirmed what James says. The Current GQ Size contains the size when conducting the listing. Current Surveys also updates the MAF in the same way as ACS.

Commented [JEZ(F8]: Add nursing home data from CMS. MEPS data?

Additional sources available for college housing GQs include data collected via web-scraping and data from the Integrated Postsecondary Education Data System (IPEDS). These variables are available at the facility level but not for individual MAFIDs.

Commented [JEZ(F9]: Is this only for 501s?

Commented [JEZ(F10R9]: I'm not sure now how we will use these data. It seems like they are most useful for determining vacant or delete status. I don't know ifw e c

Commented [TLK(F11R9]: That's right. Unless they can get pop counts soon, we won't be using the web-scraping data.

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) "**capacity utilization**"—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Additional facility-level variables may become available as research continues.

Table 4: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Data Editing

The HB edits employed to detect implausible pop counts will also be used to determine which resolved GQs contribute to the donor pool for imputation. While the most extreme outliers are flagged for imputation, for less extreme outliers, we will accept the reported values, but keep those GQs from contributing to the imputation. We compared reported pop counts with four auxiliary counts to flag outliers

Commented [JEZ(F12]: I know this is what we use, but consider editing to avoid confusion with hot deck methods. In econ we used 'imputation base' to refer to the cases that contribute to the ratios.

Commented [TLK(F13R12]: Good point. I agree with changing or droppiong the "donor pool" phrase. "Imputaiton base" avoids confusion with the hot deck, but it is a little bit confusing because it has the word "imputation" in it. I'll try to think of another phrase.
We could remove the text altogether, so it would be "contribute to the imputation."

- GQAC Expected Count
- GQAX Max Number of People
- Current Size
- Max Number of People

If the ratio between the reported pop count and the auxiliary pop count is determined to be an outlier, both the reported pop count and auxiliary count are removed from the models.

[Add a table showing how often we remove GQs from models]

Possible Methods

First, if a pop count is available from the NPC call operation and passes a quality review, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will use a combination of the following methods:

1. Ratio Imputation
2. Hierarchical Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

Commented [ADK(F14]: Need to look at paradata as covariates as well on the models

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 5 shows that 8,600 of the unresolved GQ can be resolved by converting the GQAC expected count to the GQ pop count using the ratio adjustment.

Table 5: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For each GQ type, we will use the resolved cases with a GQAC expected count to calculate the ratio of the reported GQ Census Day count to the GQAC expected count. We will then use this ratio to convert the GQAC expected count of the unresolved GQs into a Census Day imputed count. For example, for an unresolved College GQ with a GQAC Expected Count, the following equation would be applied:

$$Imputed\ Population\ Count = GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. Conversion factors for the four variables under consideration are shown in Table 6. Table 9 and Table 11 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 6: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities				
Juvenile Facilities				
Nursing Facilities				
Hospitals				
College Housing				
Military				
Shelters				
Group Homes				
Other				
All GQs				

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for colleges and universities (GQTYPCUR=501).

First, we will adjust the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population, GQAC Max Number of People, and Greek Housing variables.

After adjusting the college-level total room capacity to account reference year and for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Once we calculate the college-level residual, we will then allocate the population counts among the GQs in the college without GQAC Expected Count.

Modeling

A third approach would be to impute the GQ pop counts from a Poisson regression model. The dependent variable will be log of the ratio of reported GQ pop count to GQ Current Max Size. Independent variables are

- GQ Type
-

See Table 3 for a description of the covariates. It is important to note that GQ type is a fixed-effect covariate in the model. Each model will contain the same set of covariates, with the exception of the college model, which will also include an indicator for Greek Housing.

Median Imputation

If sufficient auxiliary data is not available, we will impute the pop size with the median population size of the resolved GQs within GQ type and state. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and state. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median GQ pop size in the cell.

Evaluation of Imputed Values

We will evaluate the imputation methods using 10-fold cross validation. First we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we divide the remaining resolved GQs into 10 approximatley equal sized groups.

We will build and fit our models on nine of the groups and then impute responses for the remaining group. We will use all four methods to impute as many units in the "unresolved" group as possible. We will do this ten times, each time treating a different group as unresolved.

Then, for each group, we will calculate the difference between the reported GQ pop and each of the four imputed methods. For each group, we will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value. We will then average these statistics across the 10 groups.

Some methods may only work under certain conditions. For example, the IPEDS residual method will only work for colleges. The Poissoin regression will only work when all of the necessary covariates are filled. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Appendix

Table 7: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Hidiroglou-Berthelot (HB) Edits

The Hidiroglou-Berthelot (HB) edit detects outliers based on the ratio of two variables. In calculating the HB statistic, the ratio is transformed once to ensure that outliers are identified at both tails of the HB statistic's distribution, then transformed again to account for the size of the reporting unit. This results in identifying the records whose data exhibit the most unusual differences between the numerator and denominator variables as well as those that have more impact on the totals. These data are identified as requiring analyst review, suppression from the imputation donor pool, or imputation.

For our purposes in this project, the HB statistic was calculated as follows.

First, we calculated the ratio between the reported pop count and the auxiliary pop count for each GQ with positive counts for both values.

$$R_i = {}^{x_i}\!/_{y_i}$$

$$x_i \ = Reported\ Pop\ Count$$

$$y_i = Auxiliary\ Pop\ Count$$

We then transformed the ratios in order to detect outliers at both tails of the distribution. We calculated median ratios within each GQ type.

$$S_i = \begin{cases} 1 - \dfrac{R_{med}}{R_i} & 0 < \ R_i < \ R_{med} \\ \dfrac{R_i}{R_{med}} - 1 & R_i > R_{med} \end{cases}$$

$$R_{med} = median\ R_i$$

When then scaled the transformed ratios by GQ size to calculate the HB statistic.

$$E_i = \ S_i * \ \sqrt{\{\max\left(x_i, y_i\right)\}}$$

To detect outliers, we calculated the following values.

$$D_{Q1} = max\{E_{med} - E_{Q1}, |.05 * E_{med}|\}$$

$$D_{Q3} = max\{E_{Q3} - E_{med}, |.05 * E_{med}|\}$$

$$E_{med} = the\ median\ value\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q1} = the\ first\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

$$E_{Q3} = the\ first\ quartile\ of\ the\ HB\ statistic\ within\ GQ\ type$$

The outliers follow outside the following range.

$$\{E_{med} - c_j * D_{Q1}, E_{med} + c_j * D_{Q3}\}$$

$$c_j = parameter\ that\ controls\ the\ width\ of\ the\ acceptance\ interval$$

We set three C values for each GQ type. The C values determined the bounds for review, suppress, and impute flags. We conducted a manual review by plotting x_i and y_i values to set the bounds by GQ type. Note, this review is somewhat subjective, but imitates common practice for establishment surveys.

References

Hidiroglou, M.A., and Berthelot, J.-M. (1986). ”Statistical Editing and Imputation for Periodic Business Surveys”. Survey Methodology, 12, 73-83.

Table 8: GQAC Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Table 9: GQAC Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Table 10: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Table 11: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

A. Calculate Ratios.
   a. For each MAFID, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘), then
      i. Assign **RATIOA** = GP/GQ_SIZE_EXP_PERS_CNT
      ii. Assign **RATIOB** = GP/GQ_SIZE_MAX_PERS_CNT
      iii. Assign **RATIOC** = GP/GQCURRSIZE
      iv. Assign **RATIOD** = GP/GQCURRMAXPOP

B. Create HB Parameters.
   a. Assign **GQTYPE** = first-digit of GQTYPCUR
   b. Create a file, *HBPARM*, with parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100
7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

b. Merge the values of C1, C2, and C3 onto the [RYAN'S FILE] file by merging HBPARM with [RYAN'S FILE] file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

d. For each MAFID, transform the ratio to create **SVALUE**.

i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)

e. For each MAFID, transform SVALUE to create **EVALUE**.

i. EVALUE = SVALUE * max {GP, GP/RATIO[X] }$^{0.5}$

ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GP is the numerator for all 4 ratios.

f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUES.

i. **E_Q1** = first quartile EVALUE

ii. **E_MED** = median EVALUE

iii. **E_Q3** = third quartile EVALUE

g. For each GQTYPE, define upper and lower bounds.

i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

iii. **LOWER_C1** = E_MED – C1 * D_Q1

iv. **LOWER_C2** = E_MED – C2 * D_Q1

v. **LOWER_C3** = E_MED – C3 * D_Q1

vi. **UPPER_C1** = E_MED + C1 * D_Q3

vii. **UPPER_C2** = E_MED + C2 * D_Q3

viii. **UPPER_C3** = E_MED + C3 * D_Q3

h. For each MAFID, create **FLAG[X]**.

i. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = 'R'

ii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'

iii. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and unres = 0. Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

i. If FLAG[X] = 'I' and GP > GP_10 and GP < GP_90 then set FLAG[X] = 'S'.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto [RYAN's FILE].

HB-Edits based on Expected Count:

GQ Type	Count not Flagged	Impute	Suppress	Review	No Expected Count or no GP	Total
Correctional Facilities	3,300	50	N<15	20	12,500	16,000
Juvenile Facilities	3,700	50	20	80	4,200	8,000
Nursing Facilities	19,000	250	20	50	9,200	28,500
Hospitals	1,200	30	20	20	1,500	2,800
College Housing	17,500	800	300	150	17,500	36,000
Military	950	20	N<15	30	4,000	5,000
Shelters	3,600	40	40	60	29,000	33,000
Group Homes	33,000	400	70	100	38,500	72,000
Other	8,900	150	30	40	17,000	26,000
All GQs	90,500	1,700	500	550	133,000	227,000

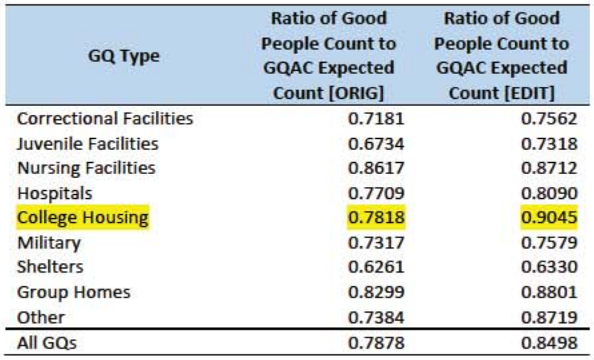
Commented [JEZ(F1]: Note, this is sum of 90,000 + 8,600 + 34,500 in Table 8.

[redacted]

Commented [JEZ(F2]: Example plot for Hospitals...(because it's easiest to see)

X = log(gp), Y = log(expected count). Ignore 'M' in legend.

I = Impute
S = Suppress
R = Review



GQ Type	Ratio of Good People Count to GQAC Expected Count [ORIG]	Ratio of Good People Count to GQAC Expected Count [EDIT]
Correctional Facilities	0.7181	0.7562
Juvenile Facilities	0.6734	0.7318
Nursing Facilities	0.8617	0.8712
Hospitals	0.7709	0.8090
College Housing	0.7818	0.9045
Military	0.7317	0.7579
Shelters	0.6261	0.6330
Group Homes	0.8299	0.8801
Other	0.7384	0.8719
All GQs	0.7878	0.8498

[redacted]

Commented [JEZ(F3]: Histogram of ratios of GP/Expected count. Bounded by [redacted]

HB-Edits based on GQAC Max Count:

GQ Type	Count not Flagged	Impute	Suppress	Review	No GQ Max or no GP	Total
Correctional Facilities	7,500	100	40	30	8,400	16,000
Juvenile Facilities	4,400	40	50	80	3,500	8,000
Nursing Facilities	20,500	150	40	40	7,500	28,500
Hospitals	1,300	40	20	30	1,400	2,800
College Housing	21,500	600	100	200	13,500	36,000
Military	1,500	40	20	40	3,500	5,000
Shelters	7,500	50	30	80	25,000	33,000
Group Homes	38,500	200	20	70	33,000	72,000
Other	10,300	200	40	30	15,500	26,000
All GQs	113,000	1,400	350	600	111,000	227,000

GQ Type	Ratio of Good People Count to GQAC Max Number of People [ORIG]	Ratio of Good People Count to GQAC Max Number of People [EDIT]
Correctional Facilities	0.4332	0.6651
Juvenile Facilities	0.2974	0.5853
Nursing Facilities	0.6603	0.7453
Hospitals	0.6391	0.7049
College Housing	0.5492	0.7910
Military	0.2290	0.6914
Shelters	0.5325	0.6152
Group Homes	0.5009	0.7926
Other	0.3783	0.6037
All GQs	0.5057	0.7308

HB-Edits based on Current Size:

GQ Type	Count not Flagged	Impute	Suppress	Review	No Current Size or no GP	Total
Correctional Facilities	7,000	150	100	90	8,700	16,000
Juvenile Facilities	3,500	20	50	60	4,400	8,000
Nursing Facilities	15,500	250	30	50	12,500	28,500
Hospitals	850	30	20	30	1,800	2,800
College Housing	17,000	450	100	90	18,500	36,000
Military	1,400	40	20	50	3,500	5,000
Shelters	4,900	40	40	100	27,500	33,000
Group Homes	29,000	250	150	80	42,000	72,000
Other	3,100	80	40	80	22,500	26,000
All GQs	82,500	1,400	550	650	142,000	227,000

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

GQ Type	Ratio of Good People Count to Current GQ Size [ORIG]	Ratio of Good People Count to Current GQ Size [EDIT]
Correctional Facilities	0.9174	0.8691
Juvenile Facilities	0.8369	0.7953
Nursing Facilities	0.9408	0.9378
Hospitals	1.0172	0.9049
College Housing	0.9444	0.9277
Military	0.9492	0.8673
Shelters	0.6180	0.6583
Group Homes	0.9679	0.9207
Other	0.9276	0.7899
All GQs	0.9217	0.8973

HB-Edits based on Max Number of People:

GQ Type	Count not Flagged	Impute	Suppress	Review	No Max or no GP	Total
Correctional Facilities	8,000	150	150	200	7,500	16,000
Juvenile Facilities	5,500	60	70	100	2,200	8,000
Nursing Facilities	24,500	300	100	70	3,600	28,500
Hospitals	1,800	50	30	40	850	2,800
College Housing	28,000	850	200	150	6,800	36,000
Military	2,500	40	20	70	2,400	5,000
Shelters	14,000	100	100	100	18,500	33,000
Group Homes	60,000	300	80	70	11,500	72,000
Other	6,300	150	100	150	19,000	26,000
All GQs	150,000	2,000	850	950	72,500	227,000

GQ Type	Ratio of Good People Count to Max Number of People [ORIG]	Ratio of Good People Count to Max Number of People [EDIT]
Correctional Facilities	0.4450	0.6743
Juvenile Facilities	0.3175	0.5877
Nursing Facilities	0.6591	0.7439
Hospitals	0.6385	0.7009
College Housing	0.5535	0.8030
Military	0.2914	0.5558
Shelters	0.5689	0.6702
Group Homes	0.4996	0.7978
Other	0.3597	0.6062
All GQs	0.5153	0.7336

GQ Type	Any flag = 'I	Any flag = 'S'	Any flag = 'I' or 'S'
Correctional Facilities			
Juvenile Facilities			
Nursing Facilities			
Hospitals			
College Housing			
Military			
Shelters			
Group Homes			
Other			
All GQs			

Commented [JEZ(F5]: don't impute but don't use in the models

Commented [JEZ(F4]: impute

Commented [JEZ(F6]: remove from models

1. Hidiroglou-Berthelot Edits

Commented [JEZ(F7]: From Service Annual Survey documentation.

1.1. Background

The Hidiroglou-Berthelot (HB) edit detects outliers based on the ratio of two variables. In calculating the HB statistic, the ratio is transformed once to ensure that outliers are identified at both tails of the HB statistic's distribution, then transformed again to account for the size and sampling weight of the reporting unit. This results in identifying the records whose data exhibit the most unusual differences between the numerator and denominator variables as well as those that have more impact on the estimates. These data are identified as requiring analyst review, suppression from the imputation base (the imputation base is described in Section 5), or imputation. For the annual surveys, the HB edit is performed on tabulation units by NAICS recode, which is defined in Attachment 2.

1.2. Calculation of the HB Statistic

First, calculate the ratio of interest, R_i, for the ith tabulation unit:

$$R_i = {}^{x_i}\!/_{y_i}$$, where

x_i = numerator for the ith tabulation unit as specified by the user

y_i = denominator for the ith tabulation unit as specified by the user.

Both x_i and y_i should be reported data values and not equal to zero.

Second, calculate the transformed ratio, S_i, as follows:

$$S_i = \begin{cases} 1 - \dfrac{R_{med}}{R_i} & 0 < R_i < R_{med} \\ \dfrac{R_i}{R_{med}} - 1 & R_i \geq R_{med} \end{cases}$$, where R_{med} = the median of the R_i.

Third, transform S_i as follows:

$E_i = S_i * \{\max(w_i x_i, w_i y_i)\}^u$, where

E_i = the HB statistic

w_i = the sampling weight

u = the size parameter ($0 \leq u \leq 1$), which we typically set to 0.5

Note that, in StEPS, there is an option to apply R_{med} to y in the second term of the HB statistic. This is done in an attempt to account for the difference in size between the two variables. Additional research is needed to determine if there is a better alternative to account for the difference, say, maybe by lowering the value of the u parameter when x and y are different variables for the same time period.

To use the HB statistic to detect outliers, calculate the following:

$$D_{Q1} = \max\{E_{med} - E_{Q1}, |A * E_{med}|\}$$

$D_{Q3} = \max\{E_{Q3} - E_{med}, |A * E_{med}|\}$, where

E_{med} = the median value of the HB statistic,

E_{Q1} = the first quartile of the HB statistic,

E_{Q3} = the third quartile of the HB statistic,

A = .05.

The outliers then fall outside this range:

$$\{E_{med} - c_j * D_{Q1}, E_{med} + c_j * D_{Q3}\}$$

c_j=the constant that determines the width of the acceptance interval

15 paths

Here is the order of imputation if we were to do so without respect to GQTYP.

Trial Ordering
(501 = CES Method on 501s) – need to give CES production dataset.
201 = impute from poisson model
101 = impute from product of GQAC expected count and GQAC expected ratio derived from GQ Type and State-Level
102 = impute from product of GQAC expected count and GQAC expected ratio derived from GQ Type
103 = impute from product of GQAC expected count and GQAC expected ratio derived from Nation
104 = impute from product of GQAC max count and GQAC max ratio derived from GQ Type and State-Level
105 = impute from product of GQAC max count and GQAC max ratio derived from GQ Type
106 = impute from product of GQAC max count and GQAC max ratio derived from Nation
107 = impute from product of Current Surveys count and Current Surveys count ratio derived from GQ Type and State-Level
108 = impute from product of Current Surveys count and Current Surveys count ratio derived from GQ Type
109 = impute from product of Current Surveys count and Current Surveys count ratio derived from Nation
110 = impute from product of Current Surveys max and Current Surveys max ratio derived from GQ Type and State-Level
111 = impute from product of Current Surveys max and Current Surveys max ratio derived from GQ Type
112 = impute from product of Current Surveys max and Current Surveys max ratio derived from Nation
401 = impute from median derived from GQ Type and State-Level
402 = impute from median derived from GQ Type
403 = impute from median derived from nation

Andrew Keller
December 24, 2020

GQ Imputation Runs

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

CODE	VALUE
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

```
Variables
GQTYPCUR: GQ Type (see above)
gqres0: Number of Resolved GQs
gpfinal0: Pop Count in Resolved GQs
gqres1: Number of Unresolved GQs
gpfinal1: Pop Count in Unresolved GQs
avgsize0: Avg GQ Size of Resolved GQs
avgsize1: Avg GQ Size of Unesolved GQs
```

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Current Production

```
                         Obs    _TYPE_    gqres0    gpfinal0    gqres1    gpfinal1    avgsize0    avgsize1

                          1        0      228000     7946000     39000     3051000      34.81       78.57
                                          ======    ========    ======    ========
                                          228000     7946000     39000     3051000


 Obs       GQTYPCUR   gqres0     gpfinal0     gqres1     gpfinal1     avgsize0   avgsize1

        1                N<15         100          .            .       12.11         .
        2        101     1100       89500       1200        46500       79.57       39.68
        3        102      250      153000       N<15         1300      667.50      150.00
        4        103     9500     1107000        350       308000      116.30      941.20
        5        104     3400      480000        250        33000      142.90      134.60
        6        105     1000       65000        250        19000       65.89       82.52
        7        106       20        1500       N<15         1600       70.29      113.40
        8        201     3800       32000        950        13500        8.35       14.22
        9        202     2000       26500        450         7600       13.02       16.91
       10        203     1400       25500        400         9600       17.75       24.39
       11        301    28000     1609000       1900       124000       57.74       63.61
       12        401      900       36000        300        21500       40.01       74.69
       13        402      300        9000        150         8700       28.40       55.23
       14        403      600        8600        200         8100       14.74       40.34
       15        404       20        1100         20         1100       64.82       50.00
       16        405      600        8400        150         4000       13.50       31.60
       17        501    35000     2612000       4400       933000       74.81      211.50
       18        601     4300      302000       1600       182000       70.94      117.20
       19        602      250       32500       N<15           60      132.90       20.00
       20        701     9200      207000       2300        68500       22.46       29.44
       21        702     4400      114000       1700        73000       26.11       42.54
       22        704      600        8300        150         6500       13.90       41.69
       23        706    34000      174000       3900        50000        5.15       12.86
       24        801    58000      506000       6800       287000        8.76       42.22
       25        802     9600      150000       1800       170000       15.58       94.30
       26        900      350        2400       N<15           80        6.68       15.20
       27        901     8100       73500       3900       420000        9.08      107.70
       28        903       60         300         30          750        5.19       27.96
       29        904    10000       95500       2800       209000        9.35       74.17
       30        999     1900       15500       2900        44000        8.06       15.14
                        ====        =====       ====
                       228000     7946000      39000      3051000
```

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

10/90 Truncation with A,B,C,D Order – Does not Include CES 501 Part

```
                       Obs    _TYPE_    gqres0    gpfinal0    gqres1    gpfinal1    avgsize0    avgsize1

                        1        0      228000     7946000     39000     1164000     34.81       29.97
                                        ======    ========    ======    ========
                                        228000     7946000     39000     1164000

 Obs      GQTYPCUR  gqres0        gpfinal0       gqres1       gpfinal1       avgsize0   avgsize1

      1                    N<15            100            .              .      12.11          .
      2        101         1100          89500         1200          46500      79.57      39.68
      3        102          250         153000         N<15            100     667.50      20.00
      4        103         9500        1107000          350          48500     116.30     147.80
      5        104         3400         480000          250          27000     142.90     110.40
      6        105         1000          65000          250          14500      65.89      61.82
      7        106           20           1500         N<15            950      70.29      68.07
      8        201         3800          32000          950           9600       8.35      10.07
      9        202         2000          26500          450           6600      13.02      14.68
     10        203         1400          25500          400           6900      17.75      17.57
     11        301        28000        1609000         1900         112000      57.74      57.53
     12        401          900          36000          300          17000      40.01      58.59
     13        402          300           9000          150           5200      28.40      33.13
     14        403          600           8600          200           2500      14.74      12.64
     15        404           20           1100           20           1200      64.82      53.91
     16        405          600           8400          150           2300      13.50      18.38
     17        501        35000        2612000         4400         348000      74.81      78.91
     18        601         4300         302000         1600          94000      70.94      60.56
     19        602          250          32500         N<15            100     132.90      30.00
     20        701         9200         207000         2300          60500      22.46      26.02
     21        702         4400         114000         1700          67500      26.11      39.41
     22        704          600           8300          150           4700      13.90      29.97
     23        706        34000         174000         3900          44500       5.15      11.41
     24        801        58000         506000         6800          56500       8.76       8.34
     25        802         9600         150000         1800          30500      15.58      16.83
     26        900          350           2400         N<15             70       6.68      14.80
     27        901         8100          73500         3900          39500       9.08      10.14
     28        903           60            300           30            450       5.19      15.74
     29        904        10000          95500         2800          87500       9.35      31.12
     30        999         1900          15500         2900          29000       8.06      10.02
                         ======         ======       ======         ======
                         228000        7946000        39000        1164000
```

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

IMP_FLAG = Imputation Flag

101 = impute from product of GQAC expected count and GQAC expected ratio derived from GQ Type and State-Level
102 = impute from product of GQAC expected count and GQAC expected ratio derived from GQ Type
103 = impute from product of GQAC expected count and GQAC expected ratio derived from Nation
104 = impute from product of GQAC max count and GQAC max ratio derived from GQ Type and State-Level
105 = impute from product of GQAC max count and GQAC max ratio derived from GQ Type
106 = impute from product of GQAC max count and GQAC max ratio derived from Nation
107 = impute from product of Current Surveys count and Current Surveys count ratio derived from GQ Type and State-Level
108 = impute from product of Current Surveys count and Current Surveys count ratio derived from GQ Type
109 = impute from product of Current Surveys count and Current Surveys count ratio derived from Nation
110 = impute from product of Current Surveys max and Current Surveys max ratio derived from GQ Type and State-Level
111 = impute from product of Current Surveys max and Current Surveys max ratio derived from GQ Type
112 = impute from product of Current Surveys max and Current Surveys max ratio derived from Nation
401 = impute from median derived from GQ Type and State-Level
402 = impute from median derived from GQ Type
403 = impute from median derived from nation

```
                                         Cumulative    Cumulative
IMP_FLAG     Frequency     Percent        Frequency       Percent
-------------------------------------------------------------
     101          7100       18.25            7100         18.25
     102            40        0.09            7100         18.34
     104          5400       13.78           12500         32.12
     105            30        0.09           12500         32.21
     106          N<15         (D)           12500         32.21
     107          5300       13.62           18000         45.83
     108           150        0.38           18000         46.21
     110          8700       22.45           26500         68.66
     111           350        0.85           27000         69.5
     112          N<15         (D)           27000         69.5
     401          6000       15.32           33000         84.84
     402          2700        6.87           35500         91.71
     403          3200        8.29           39000        100.00
```

Example: 18.25% of cases are imputed from product of GQAC expected count and GQAC expected ratio derived from GQ Type and State-Level

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

```
Obs  IMP_FLAG  _TYPE_  _FREQ_  smoothdowna  smoothupa  smoothdownb  smoothupb  smoothdownc  smoothupc  smoothdownd  smoothupd

  1     101       1      7100     0.0821      0.2162      0.0000      0.0000      0.0000     0.00000      0.0000      0.0000
  2     102       1        40     0.2571      0.1143      0.0000      0.0000      0.0000     0.00000      0.0000      0.0000
  3     104       1      5400     0.0000      0.0000      0.1559      0.0921      0.0000     0.00000      0.0000      0.0000
  4     105       1        30     0.0000      0.0000      0.2059      0.2059      0.0000     0.00000      0.0000      0.0000
  5     106       1      N<15        (D)         (D)         (D)         (D)         (D)         (D)         (D)         (D)
  6     107       1      5300     0.0000      0.0000      0.0000      0.0000      0.1131     0.09416      0.0000      0.0000
  7     108       1       150     0.0000      0.0000      0.0000      0.0000      0.5986     0.10200      0.0000      0.0000
  8     110       1      8700     0.0000      0.0000      0.0000      0.0000      0.0000     0.00000      0.1405      0.1199
  9     111       1       350     0.0000      0.0000      0.0000      0.0000      0.0000     0.00000      0.0939      0.0576
 10     112       1      N<15         (D)         (D)         (D)        (D)         (D)         (D)         (D)         (D)
 11     401       1      6000     0.0000      0.0000      0.0000      0.0000      0.0000     0.00000      0.0000      0.0000
 12     402       1      2700     0.0000      0.0000      0.0000      0.0000      0.0000     0.00000      0.0000      0.0000
 13     403       1      3200     0.0000      0.0000      0.0000      0.0000      0.0000     0.00000      0.0000      0.0000
                       ======
                        39000
```

8.2% of cases are imputed from product of GQAC expected count and GQAC expected ratio derived from GQ Type and State-Level have had their 2020 GQAC Expected Count rounded to the 10th percentile of resolved cases.

21.6% of cases are imputed from product of GQAC expected count and GQAC expected ratio derived from GQ Type and State-Level have had their 2020 GQAC Expected Count rounded to the 90th percentile of resolved cases.

1,200 cases where we impute more than the provided count.
800 cases where we impute less than the provided count.

Applying CES method to 501 cases where DSSD was using median

```
                      imp_                                     imp_
Obs    IMP_FLAG     flag_ces     _TYPE_     _FREQ_     imp_gp     gp_ces

 1        401          301          3           20       1900        450
 2        401          302          3         N<15        (D)        (D)
 3        401          304          3          700      67000      31000
 4        401          305          3           50       4400       1400
 5        401          306          3         N<15        (D)        (D)
 6        401          307          3         N<15        (D)        (D)
 7        401          308          3          100      10000       1200
                                                       ======     ======
                                                        85000      35000
```

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

10/90 Truncation with C,A,B,D Order

```
                     Obs    _TYPE_    gqres0    gpfinal0    gqres1    gpfinal1    avgsize0    avgsize1

                      1        0      228000     7946000     39000     1130000     34.81       29.11
                                      ======    ========    ======    ========
                                      228000     7946000     39000     1130000

Obs       GQTYPCUR   gqres0     gpfinal0     gqres1     gpfinal1    avgsize0   avgsize1

      1                N<15          100          .            .       12.11          .
      2       101      1100        89500       1200        46500       79.57      39.68
      3       102       250       153000       N<15          100       667.5         20
      4       103      9500      1107000        350        43500       116.3        133
      5       104      3400       480000        250        20000       142.9      81.81
      6       105      1000        65000        250        14000       65.89      60.91
      7       106        20         1500       N<15          900       70.29      65.79
      8       201      3800        32000        950         9600       8.351      10.04
      9       202      2000        26500        450         6900       13.02       15.3
     10       203      1400        25500        400         7000       17.75      17.83
     11       301     28000      1609000       1900       109000       57.74      55.92
     12       401       900        36000        300        16000       40.01      55.76
     13       402       300         9000        150         5000        28.4      32.12
     14       403       600         8600        200         2700       14.74      13.59
     15       404        20         1100         20         1200       64.82       52.5
     16       405       600         8400        150         2100        13.5      17.01
     17       501     35000      2612000       4400       336000       74.81       76.1
     18       601      4300       302000       1600        91500       70.94      58.92
     19       602       250        32500       N<15          100       132.9         30
     20       701      9200       207000       2300        60500       22.46      26.05
     21       702      4400       114000       1700        69500       26.11      40.56
     22       704       600         8300        150         4600        13.9      29.46
     23       706     34000       174000       3900        44500       5.149      11.41
     24       801     58000       506000       6800        54500       8.764      8.007
     25       802      9600       150000       1800        29500       15.58      16.29
     26       900       350         2400       N<15           70       6.675       14.8
     27       901      8100        73500       3900        38000        9.08      9.763
     28       903        60          300         30          450       5.193      17.44
     29       904     10000        95500       2800        87500       9.346      31.12
     30       999      1900        15500       2900        29000       8.062      10.04
                     ======       ======     ======       ======
                     228000      7946000      39000      1130000
```

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Issue #1 – All in one dorm
MAFID=[redacted] is a 501 where it looks like all the dorm population (of [redacted] dorms) is at that one MAFID.

Commented [TLK(F1]: This is a classic example of a GQ with an implausible population count. We should blank the pop and impute. It's okay to hard code the MAFID as unresolved.
I'm guessing POP would highlight this GQ anyway.

It has GQCURRMAXPOP=[redacted] [redacted] Hence, we are not imputing.

All the other dorms [redacted]) are [redacted], but none of the population can be distributed there since it is all in the one dorm. Can we make a rule to blank out the [redacted] MAFID?

Issue #2 – Can we stop using unrelated ratio to blank out a case?
MAFID=[redacted] is a 901 where it violates a ratio C because GQCURRSIZE=[redacted] and GQ_INITIAL_POP=[redacted]. Hence, we make it an impute.

However, we have a GQ_SIZE_EXP_PERS_CNT=[redacted] which looks pretty good in comparison to the GQ_INITIAL_POP=[redacted]. Blanking it out, we then apply a top code on the GQ_SIZE_EXP_PERS_CNT=[redacted] to EXP_PERS_TRUNC=[redacted]. This causes us to impute a count of [redacted] which had nothing to do with the original issue on the ratio of the current surveys size.

Can we only blank out and impute if the aux variable would have to be used? For example, we only blank out and impute if flagC is a bad ratio **AND** GQ_SIZE_EXP_PERS_CNT and GQ_SIZE_MAX_PERS_CNT are blank?

Commented [TKW(F2]: This makes sense to me. This is a form of error localization. We're essentially saying we have confidence in the GP/GQ_SIZE_EXP_PERS_CNT ratio and no confidence in the GP/GQCURRIZE ratio. So the error is probably in GQCURRSIZE, not GP.

Juli – If you think this is a good idea - can you integrate this logic into your code to get a new research flag file? I was using occimp_hb_dec2320.sas7bdat before.

Issue #1 – All in one dorm
MAFID=[redacted] is a 501 where it looks like all the dorm population (of [redacted] dorms) is at that one MAFID.

Commented [TLK(F1]: This is a classic example of a GQ with an implausible population count. We should blank the pop and impute. It's okay to hard code the MAFID as unresolved.
I'm guessing POP would highlight this GQ anyway.

It has GQCURRMAXPOP=[redacted] [redacted]. Hence, we are not imputing.

All the other dorms [redacted] are [redacted], but none of the population can be distributed there since it is all in the one dorm. Can we make a rule to blank out the [redacted] MAFID?

Issue #2 – Can we stop using unrelated ratio to blank out a case?
MAFID=[redacted] is a 901 where it violates a ratio C because GQCURRSIZE=[redacted] and GQ_INITIAL_POP=[redacted]. Hence, we make it an impute.

However, we have a GQ_SIZE_EXP_PERS_CNT=[redacted] which looks pretty good in comparison to the GQ_INITIAL_POP=[redacted]. Blanking it out, we then apply a top code on the GQ_SIZE_EXP_PERS_CNT=[redacted] to EXP_PERS_TRUNC=[redacted]. This causes us to impute a count of [redacted] which had nothing to do with the original issue on the ratio of the current surveys size.

Can we only blank out and impute if the aux variable would have to be used? For example, we only blank out and impute if flagC is a bad ratio **AND** GQ_SIZE_EXP_PERS_CNT and GQ_SIZE_MAX_PERS_CNT are blank?

Commented [TKW(F2]: This makes sense to me. This is a form of error localization. We're essentially saying we have confidence in the GP/GQ_SIZE_EXP_PERS_CNT ratio and no confidence in the GP/GQCURRIZE ratio. So the error is probably in GQCURRSIZE, not GP.

Juli – If you think this is a good idea - can you integrate this logic into your code to get a new research flag file? I was using occimp_hb_dec2320.sas7bdat before.

New logic:

```
a. For each MAFID, make the following updates:
   i. If FLAGA = ' ' and FLAGB = 'I' then:
      1. Set FLAGB = 'S'
      2. If FLAGC = 'I' then set FLAGC = 'S'
      3. If FLAGD = 'I' then set FLAGD = 'S'.
   ii. If FLAGA = ' ' and FLAGB = ' ' and FLAGC = 'I' then set FLAGC = 'S'
```

Commented [JEZ(F3]: In this case, FLAGA indicates the reported GP/Expected Pop count ratio is good. If we flag for imputation based on the other ratios, we'll use that expected count and likely get something close to the reported pop, anyhow.

Commented [JEZ(F4R3]: [redacted] cases.

Commented [JEZ(F5]: This is the case Andy found, where GQAC expected count and GQAC max count are in alignment with the reported count. The currsize is what is flagging these cases but we don't think they should require imputation.

Commented [JEZ(F6R5]: [redacted] cases.

Data Review from [redacted]

I looked at the [redacted], and I see an issue in the call data from the recent high-priority data collection exercise. I looked at this first given some GQ enumeration closeout notes I saw [redacted]. The total 501 Population is [redacted], when the on-campus population should be about [redacted] For all the GQs with call_status=complete, they vast majority have [redacted] Looking at the dorms, it seems clear this count is the total count for all dorms with a complete status. For example, I think the '[redacted]" probably houses probably only [redacted] students, but each GQ in the complex has a count of [redacted].

See accompanying excel file for more details.

Syntax for key cases

tab2010blkst==[redacted] & gqtypcur=="501" & tab2010blkcou==[redacted] & call_status==[redacted]

[redacted] has a imputed count + reported count that is too high. The issue is that is seems this college reported everyone in one dorm, but this didn't get picked up. Then the other dorms were imputed, generating a count that is too high.

The one dorm with a reported value of [redacted]. All combined, the imputed count is [redacted]. On Wikipedia, says [redacted]. Thus, it seems to me that the [redacted] figure is the count for all on-campus students.

TAB2010BLKST	TAB2010BLKCOU	TractGEOID	MAFID	FACTLNAME	GQNAME	GQTYPCUR
[redacted]	[redacted]	[redacted]	[redacted]		[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]		[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]		[redacted]	[redacted]

Old 2020 GQUnitPop	New 2020 GQUnitPop	CountImputed
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

Andy Comments

[redacted]

TAB2010BLKST	TAB2010BLKCOU	TractGEOID	MAFID	FACTLNAME	GQNAME	GQTYPCUR
[redacted]	[redacted]	[redacted]	[redacted]		[redacted]	[redacted]

Old_2020_GQUnitPop	New_2020_GQUnitPop	CountImputed
[redacted]	[redacted]	[redacted]

Meeting with Director Dillingham, Cogley, and Overholt

12-18-2020 - 2:00 Skype meeting – no slides

Census Attendees: Ron Jarmin, Enrique Lamas, Al Fontenot, John Abowd, Tori Velkoff, Christa Jones, Deb Stempowski, Michael Thieme, Pat Cantwell, and Leticia McCoy

Director Dillingham

1. Wasn't the goal to get to 90 percent for GQs?
2. Is it a patch that will handle getting the GQs into the DRF2?
3. Asked about the FSCPE and their role in 2010.

Cogley

"Steve great questions!"

1. Would you characterize this a hot deck imputation?
2. Little confused if you got an advanced contact number, that is not an imputation?
3. Would you consider this a statistical inference?
4. Do you have a range for the number?
5. So you used this in 2000, but Pat is not familiar?
6. This was not a part of the original operation plan? Or the August 2020 re-plan?
7. How do you deduplicate imputation?

Overholt

1. Deduplication is at the dorm level, correct?
2. Is there deduplication on imputed people?
3. Tend to have an overcount of college students?
4. So you have a question on the questionnaire to deal with college kids?
5. So we have cracks in the system?
6. Do we have knowledge about the number of college kids who were counted at mom and dad's

Overholt wants to understand the breadth of the issue. He would like to get eyes on the data.

Cogley -Sounds like this is a work in process. He would benefit from a follow up conversation when the answers to his questions are clear.

Dillingham – Thanked everyone for their hard work.

GQ Status	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Occupied GQ	179,000	17,000	1,900	197,000
Open on Census Day, Vacant During Visit	1,700	19,500	200	21,500
Refusal GQ	900	6,700	150	7,800
Vacant GQ	30,500	0	0	30,500
Delete GQ	7,600	0	0	7,600
Nonresidential GQ	2,500	0	0	2,500
Total	220,000	43,000	2,200	267,000

GQ Status	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Occupied GQ	177,000	17,000	3,100	197,000
Open on Census Day, Vacant During Visit	1,700	19,500	200	21,500
Refusal GQ	900	6,700	200	7,800
Vacant GQ	30,500	0	0	30,500
Delete GQ	7,600	0	0	7,600
Nonresidential GQ	2,500	0	0	2,500
Total	220,000	43,000	3,500	267,000

GQ Type	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Correctional Facilities*	13,000	2,800	150	16,000
Juvenile Facilities	6,100	1,800	60	8,000
Nursing Facilities*	25,000	3,200	450	28,500
Hospitals	1,900	800	60	2,800
College Housing*	29,500	5,500	650	36,000
Military*	3,100	1,900	40	5,000
Shelters	24,500	8,200	100	33,000
Group Homes	62,000	9,100	500	72,000
Other	16,000	9,700	200	26,000
Total	181,000	43,000	2,200	227,000

GQ Type	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Correctional Facilities*	13,000	2,800	250	16,000
Juvenile Facilities	6,100	1,800	90	8,000
Nursing Facilities*	24,500	3,200	600	28,500
Hospitals	1,900	800	70	2,800
College Housing*	29,000	5,500	1,300	36,000
Military*	3,000	1,900	70	5,000
Shelters	24,500	8,200	150	33,000
Group Homes	62,000	9,100	700	72,000
Other	16,000	9,700	300	26,000
Total	180,000	43,000	3,500	227,000

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

GQ Type	Suppressed from Models					
	GQAC Expected Count	GQAC Max Number of People	Current GQ Size	Max Number of People	Total Suppressed	Total Resolved
Correctional Facilities	N<15	20	100	100	200	13,000
Juvenile Facilities	20	40	40	60	100	6,100
Nursing Facilities	20	20	20	50	90	25,000
Hospitals	20	20	20	30	50	1,900
College Housing	250	60	70	150	500	29,500
Military	N<15	20	20	20	60	3,100
Shelters	30	N<15	30	100	150	24,500
Group Homes	60	N<15	150	50	250	62,000
Other	30	30	30	80	150	16,000
Total	500	200	450	650	1,500	181,000

GQ Type	Suppressed from Models					
	GQAC Expected Count	GQAC Max Number of People	Current GQ Size	Max Number of People	Total Suppressed	Total Resolved
Correctional Facilities	N<15	N<15	90	90	150	13,000
Juvenile Facilities	20	30	40	60	90	6,100
Nursing Facilities	20	N<15	20	40	80	24,500
Hospitals	20	N<15	20	30	50	1,900
College Housing	30	N<15	50	80	400	29,000
Military	N<15	20	20	N<15	50	3,000
Shelters	30	N<15	30	100	150	24,500
Group Homes	60	N<15	150	30	200	62,000
Other	30	20	30	70	150	16,000
Total	450	150	450	500	1,300	180,000

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Group Quarters Imputation Methodology

GQ Greek Indicator	GQ Count
Greek Indicator = 1	4,100
Greek Indicator = 0*	35,000
All 501s	39,500

* includes GQTYPE = 501s where GQ name is missing.

GQ Greek Indicator	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Greek Indicator = 1	0.8327	0.5443	0.8807	0.2618
Greek Indicator = 0*	0.7800	0.7216	0.9472	0.5749
All 501s	0.7818	0.5492	0.9444	0.5535

* includes GQTYPE = 501s where GQ name is missing.

GQ Type	Median Good People Count
Correctional Facilities	[redacted]
Juvenile Facilities	[redacted]
Nursing Facilities	[redacted]
Hospitals	[redacted]
College Housing	[redacted]
Military	[redacted]
Shelters	[redacted]
Group Homes	[redacted]
Other	[redacted]
All GQs	[redacted]

Tables as of Dec 23

GQ Status	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Occupied GQ	180,000	15,500	1,800	197,000
Open on Census Day, Vacant During Visit	5,000	16,500	200	21,500
Refusal GQ	2,500	5,100	150	7,800
Vacant GQ	30,500	0	0	30,500
Delete GQ	7,600	0	0	7,600
Nonresidential GQ	2,500	0	0	2,500
Total	228,000	37,000	2,200	267,000

Commented [JEZ(F1]: IMPUTE_NEEDED = 'N' have not been filtered out from this total.

GQ Type	Resolved	Unresolved		Total
		No Reported Pop	Implausible Pop	
Correctional Facilities*	14,000	1,800	150	16,000
Juvenile Facilities	6,200	1,700	60	8,000
Nursing Facilities*	26,500	1,500	450	28,500
Hospitals	2,000	750	50	2,800
College Housing*	31,500	3,900	650	36,000
Military*	3,500	1,500	40	5,000
Shelters	24,500	8,000	100	33,000
Group Homes	63,000	8,200	400	72,000
Other	16,500	9,500	200	26,000
Total	188,000	37,000	2,200	227,000

*denotes GQ Type is included in NPC calling operation

GQ Type	Suppressed from Models					
	GQAC Expected Count	GQAC Max Number of People	Current GQ Size	Max Number of People	Total Suppressed	Total Resolved
Correctional Facilities	30	60	150	150	250	14,000
Juvenile Facilities	40	50	40	80	150	6,200
Nursing Facilities	90	50	100	100	250	26,500
Hospitals	20	30	20	40	60	2,000
College Housing	700	400	250	600	1,100	31,500
Military	20	40	40	40	100	3,500
Shelters	50	20	40	150	200	24,500
Group Homes	200	100	200	200	450	63,000
Other	70	90	50	100	250	16,500
Total	1,200	800	850	1,500	2,800	188,000

GQAC Expected Count	Resolved	Unresolved	Total
Populated	95,000	7,200	102,000
Not Populated	93,000	32,000	125,000
Total	188,000	39,000	227,000

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

GQ Type	Ratio of Reported Count to GQAC Expected Count	Ratio of Reported Count to GQAC Max Number of People	Ratio of Reported Count to Current GQ Size	Ratio of Reported Count to Max Number of People
Correctional Facilities	0.7254	0 6463	0.8628	0.6565
Juvenile Facilities	0.7389	0.5801	0.7742	0.5800
Nursing Facilities	0.8674	0.7412	0.9335	0.7392
Hospitals	0.8032	0.7056	0.9018	0.6982
College Housing	0.8997	0.7835	0.9208	0.7928
Military	0.7763	0.7100	0.9236	0.5787
Shelters	0.6312	0.6091	0.6488	0.6659
Group Homes	0.8891	0.7715	0.9143	0.7771
Other	0.8491	0.6052	0.7856	0.6040
All GQs	0.8430	0.7213	0 8934	0.7238

GQAC Expected Count	Resolved	Unresolved	Total
Populated	95,000	7,200	102,000
Not Populated	93,000	32,000	125,000
Total	188,000	39,000	227,000

Table 1: GQAC Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	118,000	12,500	131,000
Not Populated	70,000	26,500	96,000
Total	188,000	39,000	227,000

Table 2: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	86,500	13,000	99,500
Not Populated	101,000	26,000	127,000
Total	188,000	39,000	227,000

Table 3: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	157,000	26,000	183,000
Not Populated	30,500	13,000	43,000
Total	188,000	39,000	227,000

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

GQ Type	Method 1	Method 1b	Method 1c	Method 1d	Method 2	Method 4	Method 3
103	1	2	5	3	4	6	
104	1	3	2	4	5	6	
105	2	4	3	5	1	6	
106							
201	2	3	1	4	5	6	
202	2	3	4	5	1	6	
203	2	4	5	3	1	6	
301	1	5	2	4	3	6	
401	3	4	5	2	1	6	
402	1	3	5	4	2	6	
403	3	1	5	2	4	6	
404							
405	1	4	2	5	3	6	
501	2	4	3	5	6	7	1
601	1	4	5	3	2	6	
701	2	4	3	5	1	6	
702	1	4	3	5	2	6	
704							
706	1	5	6	4	2	3	
801	1	2	5	3	4	6	
802	1	5	3	4	2	6	
901	1	3	5	4	2	6	
903							
999							
Average	1.53	3.53	3.79	3.89	2.68	5.89	

GQ Type	Method 1	Method 1b	Method 1c	Method 1d	Method 2	Method 4	Method 5
101	1					2	
102						1	
103	2	5	1	4	3	6	
104	1	5	3	4	2	6	
105	2	4	3	5	1	6	
106							
201	1	4	2	5	3	6	
202	1	3	5	4	2	6	
203	1					6	
301							
401							
402							
403							
404							
405							
501							
601							
701							
702							
704							
706							
801							
802							
901							
903							
999							
Average	1.3333333	4.2	2.8	4.4	2.2	6	

T. Kirk White
December 23, 2020
2020 Census Specification For “CES Method” of Group Quarters Imputation

[Can be inserted in section 3.D of “2020 Group Quarters Imputation Specification” document]

D. CES method: impute using a hybrid of the ratio imputes created in the previous step, a percentile method based on Greek/non-Greek status, and a facility-level residual allocation method.
   a. Ingest the file referred to as **MAFID_FRAT_SORO**
      i. On this file **FLAG_GREEK_LETTER**=1 indicates that GQ has been identified as a fraternity or sorority house. Otherwise **FLAG_GREEK_LETTER**=0.
   b. Ingest the file referred to as **UNITID_MAFID_LINKS**.
      i. When reading in **UNITID_MAFID_LINKS,** keep only the variables **MAFID, UNITID, MATCH_STEP_NUM**, and **ROOMCAP.**
      ii. Note: for records with **MATCH_STEP_NUM**=-1, **UNITID** will be missing.
      iii. Note: for records with the same value of UNITID, ROOMCAP will be the same.
   c. Merge **MAFID_FRAT_SORO** and **UNITID_MAFID_LINKS** to ***GQ_MAFID***, merging on MAFID, and keeping only records that are in ***GQ_MAFID.***
      i. Note: For records that match, this should be a 1-to-1 match (MAFID should be unique in each of the 3 datasets).
      ii. Note: only records with GQCURTYP=501 in ***GQ_MAFID*** should match to either of the other 2 datasets.
   d. Select the subset of the merged dataset from the previous step with GQCURTYP=501.
      i. NOTE: In this spec we will refer to this subset of the data as **GQ_COUNTS_ROOMCAP_GREEK**. This is only an intermediate dataset, which will be merged back to the **GQ_MAFID** dataset at the end of this section of the spec (section 5.D).
   e. Using GQ_COUNTS_ROOM_CAP_GREEK and the ratio impute variables created in section 4.A, create a temporary impute variable IMP_GP_TEMP using the hierarchy shown in the following table. If IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP_TEMP= IMP_RAT_EXP_GQ_ST and set ALREADY_IMPUTED=1. If IMP_RAT_EXP_GQ_ST is missing and IMP_RAT_EXP_GQ is not missing, assign IMP_GP_TEMP= IMP_RAT_EXP_GQ and set ALREADY_IMPUTED=1. Continue through the table until all the variables in the table have been exhausted. For any remaining MAFIDs for which a value has not been assigned to IMP_GP_TEMP, set ALREADY_IMPUTED=0;

IMP_GP_TEMP assignment hierarchy
IMP_RAT_EXP_GQ_ST
IMP_RAT_EXP_GQ
IMP_RAT_MAX_GQ_ST
IMP_RAT_MAX_GQ
IMP_RAT_CURR_GQ_ST
IMP_RAT_CURR_GQ
IMP_RAT_CURRMAX_GQ_ST
IMP_RAT_CURRMAX_GQ

f. Using only MAFIDs in **GQ_COUNTS_ROOMCAP_GREEK** with UNRES = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'), create 3 GP median variables and 3 GP maximum variables:
   i. For each UNITID-FLAG_GREEK_LETTER combination with enough MAFIDs:
      1. Calculate the median value of GP. Call this **P50_GP_UNIT_BY_GRK**
      2. Calculate the maximum value of GP. Call this **MAX_GP_UNIT_BY_GRK.**
      3. Merge the P50_GP_UNIT_BY_GRK and MAX_GP_UNIT_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on UNITID and FLAG_GREEK_LETTER.
   ii. For each BCUSTATEFP-FLAG_GREEK_LETTER combination with enough MAFIDs:
      1. Calculate the median value of GP. Call this **P50_GP_ST_BY_GRK**.
      2. Calculate the maximum value of GP. Call this **MAX_GP_ST_BY_GRK**.
      3. Merge P50_GP_ST_BY_GRK and MAX_GP_ST_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on BCUSTATEFP-FLAG_GREEK_LETTER combinations.
   iii. For each value of FLAG_GREEK_LETTER:
      1. Calculate the median value of GP. Call this **P50_GP_BY_GRK.**
      2. Calculate the maximum value of GP. Call this **MAX_GP_BY_GRK**.
      3. Merge P50_GP_BY_GRK and MAX_BP_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on FLAG_GREEK_LETTER.

g. For MAFIDs for which UNRES=1, FLAG_GREEK_LETTER=1, and ALREADY_IMPUTED=0, assign median Greek imputes to IMP_GP_TEMP and create up to 3 new impute variables using the following hierarchy:
   i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:
      1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
      2. Set ALREADY_IMPUTED=1
      3. Assign **MEDGP_GRK_UNIT**= IMP_GP_TEMP
   ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:
      1. assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
      2. set ALREADY_IMPUTED=1
      3. Assign **MEDGP_GRK_ST**= IMP_GP_TEMP
   iii. Otherwise:
      1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
      2. Set ALREADY_IMPUTED=1
      3. Assign **MEDGP_GRK**=IMP_GP_TEMP

h. Using **GQ_COUNTS_ROOMCAP_GREEK**, by UNITID, create unit-level sum variables (where a unit corresponds to a single UNITID, which corresponds to a single a university or college)
   i. Create unit-level sums (i.e., by UNITID) of GQCURRMAXPOP using only observations where flagD in ('','R'). Note: these are the "good" values of GQCURRMAXPOP. Note that for this sum, we don't care what the value of GP is, even it is a true 0. We are just trying to come up with a maximum number of people that these GQs *could* house, so that we can subtract the sum from the college-level IPEDS ROOMCAP variable. For reference later in the spec, call this sum **UNIT_MAXPOP_SUM**.

ii. Using only the GQs with unres=0 and flagD **not** in (‘’,’R’), by UNITID, create unit-level sums of GP. Call this sum **UNIT_2020POP_SUM**.

iii. Using only the GQs with unres=1 and flagD **not** in (‘’,’R’), by UNITID, create unit-level sums of IMP_GP_TEMP. Call this **UNIT_POP_IMPUTED_SUM**.

iv. Create **UNIT_CAP_SUM** = the unit-level sum of UNIT_MAXPOP_SUM, UNIT_2020POP_SUM, and UNIT_POP_IMPUTED_SUM

i. For each MAFID, calculate UNIT_RESIDUAL = ROOMCAP – UNIT_CAP_SUM (this will be the same value for MAFIDs with the same UNITID)

j. For each MAFID with UNIT_RESIDUAL<=0, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP, and create 3 new (non-Greek) median impute variables using the following hierarchy:

i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:

1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP

ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:

1. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP

iii. Otherwise:

1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

k. For each (non-missing) UNITID with UNIT_RESIDUAL>0, count the MAFIDs associated with that UNITID that have UNRES=1 and ALREADY_IMPUTED=0. Call this count UNIT_RESID_GQ_COUNT.

l. For MAFIDs with UNIT_RESIDUAL>0, UNIT_RESID_GQ_COUNT=1, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy:

i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK, then assign values to IMP_GP_TEMP using the following sub-hierarchy:

1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:

a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
b. Set ALREADY_IMPUTED=1
c. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP

2. Otherwise (i.e., if P50_GP_UNIT_BY_GRK<=0 or missing), if MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing, then:

a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
b. Set ALREADY_IMPUTED=1
c. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP

3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:

a. Assign IMP_GP_TEMP= P50_GP_BY_GRK
b. Set ALREADY_IMPUTED=1
c. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL < MAX_GP_UNIT_BY_GRK, then assign values as follows:
   1. Assign IMP_GP_TEMP=UNIT_RESIDUAL
   2. Set ALREADY_IMPUTED=1
   3. Assign **IMP_RESID_1GQ**=IMP_GP_TEMP

m. For MAFIDs with UNIT_RESIDUAL>0, UNIT_RESID_GQ_COUNT>1, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy. (NOTE: steps i.1-i.3 are the same as steps i.1-i.3 in step l above):
   i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK, then assign values to IMP_GP_TEMP using the following sub-hierarchy:
      1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:
         a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
         b. Set ALREADY_IMPUTED=1
         c. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP
      2. Otherwise (i.e., if P50_GP_UNIT_BY_GRK<=0 or missing), if MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing, then:
         a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
         b. Set ALREADY_IMPUTED=1
         c. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP
      3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:
         a. Assign IMP_GP_TEMP= P50_GP_BY_GRK
         b. Set ALREADY_IMPUTED=1
         c. Assign **MEDGP_nonGRK**=IMP_GP_TEMP
   ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL < MAX_GP_UNIT_BY_GRK, then assign values as follows:
      1. Assign IMP_GP_TEMP=UNIT_RESIDUAL/UNIT_RESID_GQ_COUNT
      2. Set ALREADY_IMPUTED=1
      3. Assign **IMP_RESID_NGQ**=IMP_GP_TEMP

n. Do a cross-tabulation of the variables UNRES and ALREADY_IMPUTED. If ALREADY_IMPUTED is always 1 when UNRES=1, then imputations have been calculated for all MAFIDS with GQCURTYP 501.

o. Keep the variables **MEDGP_GRK_UNIT, MEDGP_GRK_ST, MEDGP_GRK, MEDGP_nonGRK_UNIT, MEDGP_nonGRK_ST, MEDGP_nonGRK, IMP_RESID_1GQ**, and **IMP_RESID_NGQ.** Drop all other variables created in this section

Note for Andy and Juli: I believe these new variables can be inserted into the table in Section 5 of the main spec after the ratio imputation variables but before the percentile imputation variables (see table on next page). The imputations with IMP_FLAG=301 and 302 have the least mean bias, and should not be affected by the ratio imputes (except in the sense that I won't impute over a GP value that you've already create a ratio impute for). I'm less certain about the optimal order of the others in terms of mean bias. However, 304-306 are in the right order relative to each other. Also, 304-306 are only used if 307 or 308 produce bad or suspect values (i.e., negative or larger than any other GQ at that facility):

IMP_GP	**IMP_FLAG**
MEDGP_GRK_UNIT	301
MEDGP_GRK_ST	302
MEDGP_GRK	303
MEDGP_nonGRK_UNIT	304
MEDGP_nonGRK_ST	305
MEDGP_nonGRK	306
IMP_RESID_1GQ	307
IMP_RESID_NGQ	308

We created a **truth deck** to compare GQ count imputation methods. To the initial GQ universe consisted of 267, 000 units. The units were portioned by their status as determined by field operations (rows) and whether or not persons were enumerated at the unit (columns). Table 1 shows this breakdown.

Table 1: Input Data

GQ Status	No Good Person (GP)	Has Good Person	Total
Occupied GQ	17,000	181,000	197,000
Delete GQ	7,200	450	7,600
Nonresidential GQ	2,400	100	2,500
Vacant During Visit, Open on Census Day	19,500	1,900	21,500
Refusal GQ	6,700	1,100	7,800
Vacant GQ	29,000	1,100	30,500
Total	82,000	185,000	267,000

The imputation universe was defined by GQs that were determined to be occupied as of April 1 from field operations, but did not have any persons. These consisted of units classified as Occupied, Vacant during the GQ visit during July or August but open on Census Day, or GQs that refused to provide person information. These units are colored in blue. Because the units in need of imputation were all known to be occupied, the donor universe only included occupied units with a good person count (in red). This consisted of 181,000 units. Further analysis of these units removed cases where the population count was determined to be an outlier from auxiliary information obtained during the GQ Advanced Contact operation (expected count and maximum count) or information about the GQ size obtain from current surveys results (current survey count and maximum count). This reduced the occupied GQ universe with good person counts to 179,000. This was the universe that formed the basis of the truth deck.

The truth deck was created in two steps. The first step was to separate out tract-level and unit-level sampling. We wanted to sample tracts to ensure that some tracts had GQs without missing pop counts in the truth deck. To begin, the list of tracts with a GQ was split in half to form a tract-level sampling universe and a unit-level sampling universe. Among the tracts in the tract-level sampling universe, each tract was given a tract-ordering number, ordered by state, county, and then tract. The tracts were grouped together into ten samples by their last digit of their tract-ordering number.

The second list of tracts was the basis of unit-level sampling. Among the units in the unit-level sampling universe, each tract was given an unit-ordering number, ordered by GQ Type and descending GQ. The units were then grouped together into ten samples by their last digit.

To create the replicates then, samples from the tract-level sampling universe and the unit-level sampling universe were combined by their last digit on their tract-ordering number and unit-ordering number respectively. This created ten replicates of roughly 18 thousand GQs each. To test the different models, each model was fit on the 9 of the 10 replicates and then scored over the tenth replicate. This allowed for ten estimates over which statistics could be computed.

Definitions of 2020 Census Data Quality Metrics – DRAFT
Elena Healing and Mary Frances Zelenak, DSSD
November 18, 2020

2020 Census Data Quality Metrics: Apportionment Data Release

Target Release: January 7, 2021

Table 1:
Data Source: Final DRF2 (Target Completion 12/13/2020)
Universe: Enumerated MAFIDs
Geographies: US, States, DC, and Puerto Rico

Excluded from these metrics are counts from the Federally Affiliated Count Overseas and Enumeration of Transitory Locations because they are not included on the DRF2. However, their counts will be included in the apportionment counts.

NOTE: The definitions below refer to the final response associated with each housing unit and group quarters address, identified by the Master Address File Identification (MAFID), after post-data collection processing.

Quality Metric	Definition
Total Addresses (Count)	Number of distinct (unduplicated) addresses, identified by MAFID, for which an attempt to collect a census response was made through the 2020 Census operations. Includes housing unit and group quarters identified as occupied, vacant, or delete; excludes Transitory Locations and Federally Affiliated Count Overseas.
Self-Response Count	Number of distinct housing unit addresses for which a response was received through internet self-response; paper self-response including mail and Update Leave; or telephone self-response, known as Census Questionnaire Assistance. Includes housing units identified as occupied or vacant. Note that self-response operations did not identify deletes.
Nonresponse Followup Count	Number of distinct housing unit addresses for which a response was received through the Nonresponse Followup operations. Includes responses from a household member, proxy respondent, or enumerator observation; Administrative Records enumerations; and housing units identified as occupied, vacant, or delete. Excludes housing units for which no population count was collected. (These are included in Count Imputation.)
Other Count	Number of distinct housing unit and group quarters addresses for which a response was received through Update Enumerate, Remote Alaska operations, Coverage Improvement, and Group Quarters enumeration operations. Includes housing units and group quarters identified as occupied, vacant, or delete.
Count Imputation	Number of distinct housing unit addresses for which the population count was unknown. Includes addresses for which a status (occupied, vacant, or delete) or a population count is imputed, or both. Includes Nonresponse Followup addresses for which no population count was collected.

Percent Resolved as:	
Self-Response Delete	Self-response operations did not identify deletes.
Self-Response Vacant	Percent of total distinct addresses that were determined to be vacant, i.e., have no people living there, on April 1, 2020, during self-response operations.
Self-Response Occupied	Percent of total distinct addresses that were determined to be occupied, i.e., people were living there, on April 1, 2020, during self-response operations.
Nonresponse Followup Delete	Percent of total distinct addresses that were identified as 'Delete' for reasons such as: no longer existing, demolished, no longer a housing unit, or not able to locate, during nonresponse followup operations. Includes proxy responses and enumerator observation. Includes Administrative Records enumerations.
Nonresponse Followup Vacant	Percent of total distinct addresses that were determined to be vacant, i.e., have no people living there, on April 1, 2020, during nonresponse followup operations. Includes proxy responses and enumerator observation. Includes Administrative Records enumerations.
Nonresponse Followup Occupied	Percent of total distinct addresses that were determined to be occupied, i.e., people were living there, on April 1, 2020, during nonresponse followup operations. Includes responses from a household member or proxy respondent. Includes housing units for which only a population count was collected. Includes Administrative Records enumerations. Excludes housing units for which no population count was collected. (These are included in Count Imputation.)
Other Delete (UE, RA, GQE, etc)	Percent of total distinct addresses that were determined to no longer be a housing unit or group quarters on April 1, 2020, during Update Enumerate, Update Leave, Remote Alaska, Group Quarters Enumeration operations.
Other Vacant (UE, RA, GQE, etc)	Percent of total distinct addresses that were determined to be vacant, i.e., have no people living there, on April 1, 2020, during Update Enumerate, Remote Alaska, Coverage Improvement, and Group Quarters enumeration operations.
Other Occupied (UE, RA, GQE, etc)	Percent of total distinct addresses that were determined to be occupied, i.e., people were living there, on April 1, 2020, during Update Enumerate, Remote Alaska operations, Coverage Improvement, and Group Quarters enumeration operations.
Unresolved (went to Count Imputation)	Percent of total distinct addresses for which the population count was unknown. Includes housing units for which the final status from enumeration efforts was unresolved for reasons such as 'occupied without a population count', 'unresolved, address exists, but unknown occupancy status', or otherwise unresolved. Includes addresses for which a status (occupied, vacant, or delete) or a population count is imputed, or both.If determined to be occupied, a corresponding population count was assigned during imputation. Includes

	housing units from nonresponse followup for which no population count was collected.
Percent Resolved as:	
Self-Response	Number of distinct housing unit addresses for which a response was received through internet self-response; paper self-response including mail and Update Leave; or telephone self-response, known as Census Questionnaire Assistance. Includes housing units identified as occupied, vacant, or delete.
Internet	Of the addresses enumerated by self-response, the percent for which a response was received through internet self-response.
Paper	Of the addresses enumerated by self-response, the percent for which a response was received through paper self-response including mail and Update Leave.
Telephone	Of the addresses enumerated by self-response, the percent for which a response was received through telephone self-response, Census Questionnaire Assistance.
All Nonresponse Followup Activity	Number of distinct housing unit addresses for which a response was received through the Nonresponse Followup operations. Includes responses from a household member, proxy respondent, or enumerator observation; Administrative Records enumerations; and housing units identified as occupied, vacant, or delete.
Household Interview	Of the addresses enumerated by nonresponse followup, the percent for which a response was received from a household member.
Proxy	Of the addresses enumerated by nonresponse followup, the percent for which a response was received from a proxy respondent.
Administrative Records	Of the addresses enumerated by nonresponse followup, the percent for which administrative records were used for enumeration. Administrative records were used after unsuccessful attempts to enumerate the housing units through self-response, nonresponse followup, or other census operations.
Delete	Of the addresses enumerated by Administrative Records, the percent that were determined to no longer be a housing unit on April 1, 2020. Reasons for this status include records of the housing unit being demolished and conversion of the housing unit to a business or office.
Vacant	Of the addresses enumerated by Administrative Records, the percent that were determined to be vacant, i.e., have no people living there, on April 1, 2020.
Occupied	Of the addresses enumerated by Administrative Records, the percent that were determined to be occupied, i.e., people were living there, on April 1, 2020, based on administrative records.
Percent Resolved as:	
Nonresponse Followup POP Count Only	Number of addresses enumerated by nonresponse followup for which only a population count was collected. This is a subset of the nonresponse followup occupied addresses.

(Subset of NRFU Occupied)	
Household Interview	Of the addresses enumerated by nonresponse followup with only a population count, the percent for which a response was received from a household member.
Proxy Interview	Of the addresses enumerated by nonresponse followup with only a population count, the percent for which a response was received from a proxy respondent.
Alternate Metrics for POP Count Only	
Percent of Household Interview	*Of the nonresponse followup addresses with a response from a household member, the percent with only a population count.*
Percent of Proxy Interview	*Of the nonresponse followup addresses with a response from a proxy, the percent with only a population count.*

Andrew Keller, Julianne Zamora, Tim Kennel
December 21, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input File:

1. /p2020_drfrv/t_king0345/gq_mafid_cnts_121920_geo_cdl.sas7bdat
2. CES 501 results
3. CES 301 results

Output File: DSSD GQ Imputation File

Section 1: Defining the Unresolved Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingesting the input File, we must initially determine what is eligible for imputation. For the cases not eligible for imputation, we assign three variables to determine this universe:
    a. **gp_initial** = This is the count of good persons in the GQ prior to imputation (0,1,….)
    b. **gpy_initial** = This indicates whether the GQ has any good persons (0/1)
    c. **unres_initial** = This indicates whether the GQ is unresolved and eligible to be imputed a positive pop count. (0/1)

```
12/21/2020
TO BEGIN: SKIP ALL the LOGIC in this Section (A) and use this:

     if GP>0 and GP_PSA>0 then GP=GP_PSA;
     else if GP>0 and GP_PSA=. then GP=GP;
     else if GP=. and ddp in (0,.) then GP=max(CDLPER,GEO_POP_COUNT);

   if gp > 0 then gpy = 1; else gpy = 0;

 unres1 = 0;
if FOCS_ER_CB_CODE  in ('','O','R') and gpy = 0 then unres1 = 1;

unres2 = unres1;
if IMPUTE_NEEDED = 'N' then unres2 = 0;
```

```
unres=unres2;
```

1. **To determine the GQ status: start with FOCS_ER_CB_CODE**

2. **To determine the GQ has good persons (and the GQ count), I use the gp value, but I overwrite with this logic.**

   if gp_psa > 0 then gp_initial = gp_psa
   if gp_initial = . and ddp = (0,.) then gp_initial = cdlper
   if gp_initial > 0 then gpy_initial = 1; else gpy_initial = 0;

3. **To determine the unresolved cases:**

   unres_initial = 0;
   if FOCS_ER_CB_CODE in ('','O','R') and gpy_initial = 0 then unres_initial = 1;
   ADK: GOTTA ADD HOW WE TAKE OUT IMPUTE_NEEDED cases and give 0 pop count if necessary

B. JEZ After making initial determinations on what is eligible for imputation, we must removed outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.
   a. **GP** = This is the count of good persons in the GQ prior to imputation (0,1,....)
   b. **GPY** = This indicates whether the GQ has any good persons (0/1)
   c. **UNRES** = This indicates whether the GQ is unresolved and eligble to be imputed an positive pop count. (0/1)

Section 2: Defining the Unresolved Cases Eligible for GQ Size Imputation
This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values for when GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.
   a. We will create 3 ratios for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. If GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = ''
      i. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
      ii. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
      iii. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
      iv. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
      v. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
      vi. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
      vii. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
      viii. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
      ix. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**

Commented [JEZ(F1]: Why are these conditions on the creation of the ratios?

I would just calculate the ratios first, and then use the conditions you have to decide when to use them.

I don't understand this sub-setting. I would subset the universe for each ratio separately.

EXPRATIO = sum(GP)/sum (GQ_SIZE_EXP_PERS_CNT) where unres = '0' and FOCS_ER_CB_CODE ' ' and flagA in (' ', 'R')

MAXRATIO = sum(GP)/sum (GQ_SIZE_MAX_PERS_CNT) where unres = '0' and FOCS_ER_CB_CODE = ' ' and flagB in (' ','R')

Etc. It will be easier to code this way and it will make maximum use of the reported data.

I think you only need three sets of ratios for each of the four variables, so only 12 applicable factors for each GQTYPCUR. I think the conditions on which variables are populated only matter for the business rules at the end.

Commented [JEZ(F2R1]: I added a table at the end of the document to show what I think we should do, how we could spec out the 12 ratios.

x. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
xi. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
xii. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
xiii. Sum the GP and GQCURRSIZE value **for the nation.**
xiv. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
xv. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
xvi. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
xvii. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
xviii. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
xix. Sum the GP and GQCURRMAXPOP value **for the nation.**
xx. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
xxi. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
xxii. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
xxiii. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
xxiv. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

B. Assign Ratio-Adjustment Values for when at least one of GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT is greater than 0, but they all are not (since it is covered in the case above.

a. We will create 3 ratios for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. **<u>If it is not true that all</u>** GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0, but GQ_SIZE_EXP_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

i. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
ii. Assign **EXPRATIO1** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
iii. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
iv. Assign **EXPRATIO1_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
v. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
vi. Assign **EXPRATIO1_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

b. We will create 3 ratios for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. **<u>If it is not true that all</u>** GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0, but GQ_SIZE_MAX_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

i. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
ii. Assign **MAXRATIO1** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
iii. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**

iv. Assign **MAXRATIO1_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID

v. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**

vi. Assign **MAXRATIO1_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

c. We will create 3 ratios for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. **<u>If it is not true that all</u>** GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0, but GQCURRSIZE > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

i. Sum the GP and GQCURRSIZE value for the nation.

ii. Assign **CURRSIZERATIO1** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

iii. Sum the GP and GQCURRSIZE value for each GQTYPCUR value.

iv. Assign **CURRSIZERATIO1_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID

v. Sum the GP and GQCURRSIZE value for each combination of GQTYPCUR and BCUSTATEFP value.

vi. Assign **CURRSIZERATIO1_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

d. We will create 3 ratios for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. **<u>If it is not true that all</u>** GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0, but GQCURRMAXPOP > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

i. Sum the GP and GQCURRMAXPOP value **for the nation.**

ii. Assign **CURRMAXRATIO1** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

iii. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**

iv. Assign **CURRMAXRATIO1_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID

v. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**

vi. Assign **CURRMAXRATIO1_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

C. Assign Good Person Percentile counts for when GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **<u>all</u>** greater than 0.

a. We will create 3 Good Person Percentile counts for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. Do this if GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**

2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

D. Assign Good Person Percentile counts for when at least one of GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT is greater than 0, but they all are not (since it is covered in the case above.

  a. We will create 3 Good Person Percentile counts for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. Do this **<u>if it is not true that all</u>** GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0, but at least one of the four is greater than 0 and unres = 0 and FOCS_ER_CB_CODE = ''

    i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP1.**

    ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP1_GQ.**

    iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP1_GQ_ST.**

      1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP1_GQ_ST.**
      2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP1_GQ_ST.**
      3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP1_GQ_ST.**

E. Assign Good Person Percentile counts when GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are all 0.

  a. We will create 3 Good Person Percentile counts for each variable, one for the national value, one for the GQTYPCUR combination, and one for the GQTYPCUR and BCUSTATEFP combination. Do this **<u>if all</u>** GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are 0 and unres = 0 and FOCS_ER_CB_CODE = ''

    i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP0.**

    ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP0_GQ.**

    iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP0_GQ_ST.**

      1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP0_GQ_ST.**
      2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP0_GQ_ST.**

3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP0_GQ_ST.**

F. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **<u>all</u>** greater than 0.

   a. Define MAXPOP variable.
```
        if gqcurrmaxpop > 0 then maxpop = log(gqcurrmaxpop);
        if gqcurrmaxpop = 0 then maxpop = .;
```
   b. Define the fitting universe (ratiofile) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’
   c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.
   d. Fit and score this model:
```
proc genmod data = ratiofile;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

   e. Take the ceiling function of the predicted count. Call this **poisson_count.**

G. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **<u>all</u>** greater than 0.

H. Fold in CES 501 results

I. Fold in CES 301 results

<u>Section 3: Applying Business Rules</u>

The next section assigns the imputed values. It is broken into three sections based on the auxiliary data.

- GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **<u>all</u>** greater than 0.
- at least one of GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT is greater than 0, but they all are not (since it is covered in the case above
- GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are all 0.

A. Define these variables:

a. if GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 then hasall = 1; else hasall = 0;
b. Business Rules:

```
if GQTYPCUR = '104' and CURRSIZERATIO_GQ_ST > 0 and hasall = 1 then do;
GQIMPCT = CEIL(CURRSIZERATIO_GQ_ST * GQCURRSIZE);
GQIMPPATH = 109;
end;

else if GQTYPCUR = '104' and CURRSIZERATIO_GQ_ST <= 0 and CURRSIZERATIO_GQ > 0 and hasall
= 1 then do;
GQIMPCT = CEIL(CURRSIZERATIO_GQ * GQCURRSIZE);
GQIMPPATH = 108;
end;

else if GQTYPCUR = '104' and CURRSIZERATIO_GQ <= 0 and CURRSIZERATIO > 0 and hasall = 1
then do;
GQIMPCT = CEIL(CURRSIZERATIO * GQCURRSIZE);
GQIMPPATH = 107;
end;

if GQTYPCUR = '105' and EXPRATIO_GQ_ST > 0 and hasall = 1 then do;
GQIMPCT = CEIL(EXPRATIO_GQ_ST * GQ_SIZE_EXP_PERS_CNT);
GQIMPPATH = 103;
end;

else if GQTYPCUR = '105' and EXPRATIO_GQ_ST <= 0 and EXPRATIO_GQ > 0 and hasall = 1 then
do;
GQIMPCT = CEIL(EXPRATIO_GQ * GQ_SIZE_EXP_PERS_CNT);
GQIMPPATH = 102;
end;

else if GQTYPCUR = '105' and EXPRATIO_GQ <= 0 and EXPRATIO > 0 and hasall = 1 then do;
GQIMPCT = CEIL(EXPRATIO * GQ_SIZE_EXP_PERS_CNT);
GQIMPPATH = 101;
end;
```

GQTYPCUR	**Condition (s)**	**Method**	**Flag**
104	GQCURRSIZE > 0 and CURRSIZE_RATIO_GQ_ST > 0	CEIL (CURRSIZE_RATIO_GQ_ST * GQCURRSIZE)	GQIMPPATH = 1(
104	GQCURRSIZE > 0 and GQCURRSIZE_GQ > 0	CEIL (CURRSIZE_RATIO_GQ * GQCURRSIZE)	GQIMPPATH = 1(

Commented [JEZ(F3]: You could do a table like this and write instructions that say, do the imputation by GQTYPCUR. If a MAFID meets the set of conditions, use the method to impute the value and set the flag, if not, move to the next row, etc.

RATIOS:

Create the following ratios by summing values for all IDs where unres = 0 and FOCS_ER_CB_CODE = ‘ ‘. Use the table to determine the level for the ratio and any additional conditions. For example,

$$EXPRATIO_{GQ\_ST} = \frac{\sum_i GP}{\sum_i GQ\_SIZE\_EXP\_PERS\_CNT}$$

$$where\ i\ in\ GQTYPCUR\ and\ BCUSTATEFP\ and\ FLAGA\ in\ ('\ \ ','\ R')$$

Ratio	Numerator	Denominator	Level	Condition
EXPRATIO_GQ_ST	SUM(GP)	SUM(GQ_SIZE_EXP_PERS_CNT)	GQTYPCUR*BCUSTATEFP	FLAGA in (‘ ‘,’R’)
EXPRATIO_GQ	SUM(GP)	SUM(GQ_SIZE_EXP_PERS_CNT)	GQTYPCUR	FLAGA in (‘ ‘,’R’)
EXPRATIO	SUM(GP)	SUM(GQ_SIZE_EXP_PERS_CNT)	All MAFIDs meeting conditions	FLAGA in (‘ ‘, ‘R’)
MAXRATIO_GQ_ST	SUM(GP)	SUM(GQ_SIZE_MAX_PERS_CNT)	GQTYPCUR*BCUSTATEFP	FLAGB in (‘ ‘,’R’)
…				

Andrew Keller, Julianne Zamora, Tim Kennel
December 23, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input Files:

1. /sampling/eb/kelle321/gq_mafid_cnts_121920_geo_cdl.sas7bdat
2. /sampling/share/hbparm.sas7bdat
3. CES 501 results
4. CES 301 results

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file, referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.

Rename GQ_INITIAL_STATUS to GQ_PRE_STATUS.
Rename GQ_INITIAL_UNRES to GQ_PRE_UNRES.
Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file, referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.
D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
    a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

Section 2: HB Edits

A. Calculate Ratios for editing.

a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in ('O','R',' '), then

i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT

ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT

iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE

iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP

b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.

a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR

b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100

7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

d. For each MAFID, transform the ratio to create **SVALUE**.

i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)

e. For each MAFID, transform SVALUE to create **EVALUE**.

i. EVALUE = SVALUE * max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}$^{0.5}$

ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

i. **E_Q1** = first quartile EVALUE

ii. **E_MED** = median EVALUE

iii. **E_Q3** = third quartile EVALUE

g. For each GQTYPE, define upper and lower bounds.

i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

iii. **LOWER_C1** = E_MED – C1 * D_Q1

iv. **LOWER_C2** = E_MED – C2 * D_Q1

v. **LOWER_C3** = E_MED – C3 * D_Q1

vi. **UPPER_C1** = E_MED + C1 * D_Q3

vii. **UPPER_C2** = E_MED + C2 * D_Q3

viii. **UPPER_C3** = E_MED + C3 * D_Q3

h. For each MAFID, create **FLAG[X]**.

i. If EVALUE is missing, FLAG[X] = ‘M’

ii. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = ‘R’

iii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = ‘S’

iv. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = ‘I’

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (‘ ‘,’O’,’R’) and GQ_INITIAL_UNRES = 0. Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

i. If FLAG[X] = ‘I’ and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = ‘S’.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

a. If (FLAGA = ‘I’ or FLAGB = ‘I’ or FLAGC = ‘I’ or FLAGD = ‘I’) and IMPUTE_NEEDED ne ‘N’ then

i. **GP = .**

ii. **UNRES** = 1

b. Otherwise,

i. **GP =** GQ_INITIAL_POP

ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’):

1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_EXP** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO)
2. **IMP_RAT_EXP_GQ** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ)
3. **IMP_RAT_EXP_GQ_ST** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values

   i. Calculate Ratios.
      We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):

      1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
      2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
      3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
      4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
      5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
      6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

   ii. Assign values. For each MAFID, calculate the following values:

      1. **IMP_RAT_MAX** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO)
      2. **IMP_RAT_MAX_GQ** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ)
      3. **IMPRAT_MAX_GQ_ST** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

   i. Calculate Ratios.
      We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):

      1. Sum the GP and GQCURRSIZE value **for the nation.**
      2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
      3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
      4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
      5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
      6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

   ii. Assign values. For each MAFID, calculate the following values:

      1. **IMP_RAT_CURR** = CEIL (GQCURRSIZE*CURRSIZERATIO)
      2. **IMP_RAT_CURR_GQ** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ)
      3. **IMP_RAT_CURR_GQ_ST** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

   i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (GQCURRMAXPOP*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

a. Define MAXPOP variable.

i. if GQCURRMAXPOP > 0 then **MAXPOP** = log(GQCURRMAXPOP);

ii. if GQCURRMAXPOP = 0 then **MAXPOP** = .;

b. Define the fitting universe (ratiofile) as this: FLAGA in (‘ ‘,’R’) and FLAGB in (‘ ‘,’R’) and FLAGC in (‘ ‘,’R’) and FLAGD in (‘ ‘,’R’) and unres = 0 and FOCS_ER_CB_CODE = ‘’

c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.

d. Fit and score this model:

```
proc genmod data = ratiofile;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **IMP_POISSON_COUNT.**

Commented [JEZ(F1]: Remove?

D. Fold in CES 501 results

Commented [JEZ(F2]: Residual Method

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table as follows, if IMP_POISSON_COUNT is not missing, assign IMP_GP = IMP_POISSON_COUNT and assign IMP_FLAG = 201. If IMP_POISSON_COUNT is missing, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. Continue on through the table until all MAFIDs in unres = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	**IMP_FLAG**
IMP_POISSON_COUNT	201
IMP_RAT_EXP_GQ_ST	101
IMP_RAT_EXP_GQ	102
IMP_RAT_EXP	103
IMP_RAT_MAX_GQ_ST	104
IMP_RAT_MAX_GQ	105
IMP_RAT_MAX	106
IMP_RAT_CURR_GQ_ST	107
IMP_RAT_CURR_GQ	108
‘IMP_RAT_CURR	109
IMP_RAT_CURRMAX_GQ_ST	110
IMP_RAT_CURRMAX_GQ	111
IMP_RAT_CURRMAX	112
MEDGP_GQ_ST	401
MEDGP_GQ	402
MEDGP	403

Commented [JEZ(F3]: Remove?

Section 6: Create Output File

Output GQ_MAFID, adding the following variables:

FLAGA	FLAGB	
FLAGC	FLAGD	
GP	UNRES	
EXPRATIO	EXPRATIO_GQ	EXPRATIO_GQ_ST
IMP_RAT_EXP	IMP_RAT_EXP_GQ	IMP_RAT_EXP_GQ_ST
MAXRATIO	MAXRATIO_GQ	MAXRATIO_GQ_ST
IMP_RAT_MAX	IMP_RAT_MAX_GQ	IMP_RAT_MAX_GQ_ST
CURRRATIO	CURRRATIO_GQ	CURRATIO_GQ_ST
IMP_RAT_CURR	IMP_RAT_CURR_GQ	IMP_RAT_CURR_GQ_ST
MAXCURRRATIO	MAXCURRRATIO_GQ	MAXCURRRATIO_GQ_ST
IMP_RAT_MAXCURR	IMP_RAT_MAXCURR_GQ	IMP_RAT_MAXCURR_GQ_ST
MEDGP	MEDGP_GQ	MEDGP_GQ_ST
IMP_GP	IMP_FLAG	

Name this file gq_mafid_dssd_out.sas7bdat

Andrew Keller, Julianne Zamora, Tim Kennel
December 21, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input File:

1. /p2020_drfrv/t_king0345/gq_mafid_cnts_121920_geo_cdl.sas7bdat
2. CES 501 results
3. CES 301 results

Output File: DSSD GQ Imputation File

Section 1: Defining the Unresolved Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingesting the input File, we must initially determine what is eligible for imputation. For the cases not eligible for imputation, we assign three variables to determine this universe:
   a. **gp_initial** = This is the count of good persons in the GQ prior to imputation (0,1,….)
   b. **gpy_initial** = This indicates whether the GQ has any good persons (0/1)
   c. **unres_initial** = This indicates whether the GQ is unresolved and eligible to be imputed a positive pop count. (0/1)

```
12/21/2020
TO BEGIN: SKIP ALL the LOGIC in this Section (A) and use this:

     if GP>0 and GP_PSA>0 then GP=GP_PSA;
     else if GP>0 and GP_PSA=. then GP=GP;
     else if GP=. and ddp in (0,.) then GP=max(CDLPER,GEO_POP_COUNT);

   if gp > 0 then gpy = 1; else gpy = 0;

 unres1 = 0;
if FOCS_ER_CB_CODE  in ('','O','R') and gpy = 0 then unres1 = 1;

unres2 = unres1;
if IMPUTE_NEEDED = 'N' then unres2 = 0;
```

```
unres=unres2;
```

1. **To determine the GQ status: start with FOCS_ER_CB_CODE**

2. **To determine the GQ has good persons (and the GQ count), I use the gp value, but I overwrite with this logic.**

    if gp_psa > 0 then gp_initial = gp_psa
    if gp_initial = . and ddp = (0,.) then gp_initial = cdlper
    if gp_initial > 0 then gpy_initial = 1; else gpy_initial = 0;

3. **To determine the unresolved cases:**

    unres_initial = 0;
    if FOCS_ER_CB_CODE in ('','O','R') and gpy_initial = 0 then unres_initial = 1;
    ADK: GOTTA ADD HOW WE TAKE OUT IMPUTE_NEEDED cases and give 0 pop count if necessary

B. JEZ After making initial determinations on what is eligible for imputation, we must removed outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

  a. **GP** = This is the count of good persons in the GQ prior to imputation (0,1,....)
  b. **GPY** = This indicates whether the GQ has any good persons (0/1)
  c. **UNRES** = This indicates whether the GQ is unresolved and eligble to be imputed an positive pop count. (0/1)

Section 2: Defining the Unresolved Cases Eligible for GQ Size Imputation

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

  a. We will create 3 ratios for each variable, one for the national value (**EXPRATIO)**, one for the GQTYPCUR combination (**EXPRATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):
    i. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
    ii. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
    iii. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
    iv. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
    v. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
    vi. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
  b. We will create 3 ratios for each variable, one for the national value (**MAXRATIO)**, one for the GQTYPCUR combination (**MAXRATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):

Commented [JEZ(F1]: Why are these conditions on the creation of the ratios?

I would just calculate the ratios first, and then use the conditions you have to decide when to use them.

I don't understand this sub-setting. I would subset the universe for each ratio separately.

EXPRATIO = sum(GP)/sum (GQ_SIZE_EXP_PERS_CNT) where unres = '0' and FOCS_ER_CB_CODE ' ' and flagA in (' ', 'R')

MAXRATIO = sum(GP)/sum (GQ_SIZE_MAX_PERS_CNT) where unres = '0' and FOCS_ER_CB_CODE = ' ' and flagB in (' ','R')

Etc. It will be easier to code this way and it will make maximum use of the reported data.

I think you only need three sets of ratios for each of the four variables, so only 12 applicable factors for each GQTYPCUR. I think the conditions on which variables are populated only matter for the business rules at the end.

Commented [JEZ(F2R1]: I added a table at the end of the document to show what I think we should do, how we could spec out the 12 ratios.

  i. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
  ii. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
  iii. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
  iv. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
  v. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
  vi. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

c. We will create 3 ratios for each variable, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagC in (‘’,’R’):
  i. Sum the GP and GQCURRSIZE value **for the nation.**
  ii. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
  iii. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
  iv. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
  v. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
  vi. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

d. We will create 3 ratios for each variable, one for the national value (**CURRMAXRATIO)**, one for the GQTYPCUR combination (**CURRMAXRATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagD in (‘’,’R’):
  i. Sum the GP and GQCURRMAXPOP value **for the nation.**
  ii. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
  iii. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
  iv. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
  v. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
  vi. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP)**, one for the GQTYPCUR combination (**MEDGP_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’,’R’):
  i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**
  ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**
  iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**
    1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**

2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

a. Define MAXPOP variable.

```
      if gqcurrmaxpop > 0 then maxpop = log(gqcurrmaxpop);
      if gqcurrmaxpop = 0 then maxpop = .;
```

b. Define the fitting universe (ratiofile) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.

d. Fit and score this model:

```
proc genmod data = ratiofile;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **poisson_count.**

D. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

E. Fold in CES 501 results

F. Fold in CES 301 results

Section 3: Applying Business Rules

The next section assigns the imputed values. It is broken into three sections based on the auxiliary data.

A. GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

GQTYPCUR	Condition (s)	Method	Flag

Commented [JEZ(F3]: You could do a table like this and write instructions that say, do the imputation by GQTYPCUR. If a MAFID meets the set of conditions, use the method to impute the value and set the flag, if not, move to the next row, etc.

104	GQCURRSIZE > 0 and CURRSIZE_RATIO_GQ_ST > 0	CEIL (CURRSIZE_RATIO_GQ_ST * GQCURRSIZE)	GQIMPPATH = 109
104	GQCURRSIZE > 0 and GQCURRSIZE_GQ > 0	CEIL (CURRSIZE_RATIO_GQ * GQCURRSIZE)	GQIMPPATH = 108

B. at least one of GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT is greater than 0, but they all are not (since it is covered in the case above
C. GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are all 0.

RATIOS:

Create the following ratios by summing values for all IDs where unres = 0 and FOCS_ER_CB_CODE = ‘ ‘. Use the table to determine the level for the ratio and any additional conditions. For example,

$$EXPRATIO_{GQ\_ST} = \frac{\sum_i GP}{\sum_i GQ\_SIZE\_EXP\_PERS\_CNT}$$

$$where\ i\ in\ GQTYPCUR\ and\ BCUSTATEFP\ and\ FLAGA\ in\ ('\ ','\ R')$$

Ratio	Numerator	Denominator	Level	Condition
EXPRATIO_GQ_ST	SUM(GP)	SUM(GQ_SIZE_EXP_PERS_CNT)	GQTYPCUR*BCUSTATEFP	FLAGA in (‘ ‘,’R’)
EXPRATIO_GQ	SUM(GP)	SUM(GQ_SIZE_EXP_PERS_CNT)	GQTYPCUR	FLAGA in (‘ ‘,’R’)
EXPRATIO	SUM(GP)	SUM(GQ_SIZE_EXP_PERS_CNT)	All MAFIDs meeting conditions	FLAGA in (‘ ‘, ‘R’)
MAXRATIO_GQ_ST	SUM(GP)	SUM(GQ_SIZE_MAX_PERS_CNT)	GQTYPCUR*BCUSTATEFP	FLAGB in (‘ ‘,’R’)
…				

Andrew Keller, Julianne Zamora, Tim Kennel
December 21, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input File:

1. /p2020_drfrv/t_king0345/gq_mafid_cnts_121920_geo_cdl.sas7bdat
2. CES 501 results
3. CES 301 results

Output File: DSSD GQ Imputation File

Section 1: Defining the Unresolved Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingesting the input File, we must initially determine what is eligible for imputation. For the cases not eligible for imputation, we assign three variables to determine this universe:
    a. **gp_initial** = This is the count of good persons in the GQ prior to imputation (0,1,….)
    b. **gpy_initial** = This indicates whether the GQ has any good persons (0/1)
    c. **unres_initial** = This indicates whether the GQ is unresolved and eligible to be imputed a positive pop count. (0/1)

```
12/21/2020
TO BEGIN: SKIP ALL the LOGIC in this Section (A) and use this:

     if GP>0 and GP_PSA>0 then GP=GP_PSA;
     else if GP>0 and GP_PSA=. then GP=GP;
     else if GP=. and ddp in (0,.) then GP=max(CDLPER,GEO_POP_COUNT);

   if gp > 0 then gpy = 1; else gpy = 0;

 unres1 = 0;
if FOCS_ER_CB_CODE  in ('','O','R') and gpy = 0 then unres1 = 1;

unres2 = unres1;
if IMPUTE_NEEDED = 'N' then unres2 = 0;
```

```
unres=unres2;
```

1. **To determine the GQ status: start with FOCS_ER_CB_CODE**

2. **To determine the GQ has good persons (and the GQ count), I use the gp value, but I overwrite with this logic.**

   if gp_psa > 0 then gp_initial = gp_psa
   if gp_initial = . and ddp = (0,.) then gp_initial = cdlper
   if gp_initial > 0 then gpy_initial = 1; else gpy_initial = 0;

3. **To determine the unresolved cases:**

   unres_initial = 0;
   if FOCS_ER_CB_CODE in ('','O','R') and gpy_initial = 0 then unres_initial = 1;
   ADK: GOTTA ADD HOW WE TAKE OUT IMPUTE_NEEDED cases and give 0 pop count if necessary

B. **JEZ** After making initial determinations on what is eligible for imputation, we must removed outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.
   a. **GP** = This is the count of good persons in the GQ prior to imputation (0,1,….)
   b. **GPY** = This indicates whether the GQ has any good persons (0/1)
   c. **UNRES** = This indicates whether the GQ is unresolved and eligble to be imputed an positive pop count. (0/1)

Section 2: Defining the Unresolved Cases Eligible for GQ Size Imputation

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values
   a. We will create 3 ratios for each variable, one for the national value (**EXPRATIO)**, one for the GQTYPCUR combination (**EXPRATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’):
      i. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
      ii. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
      iii. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
      iv. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
      v. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
      vi. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
   b. We will create 3 ratios for each variable, one for the national value (**MAXRATIO)**, one for the GQTYPCUR combination (**MAXRATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagB in (‘’,’R’):

Commented [JEZ(F1]: Why are these conditions on the creation of the ratios?

I would just calculate the ratios first, and then use the conditions you have to decide when to use them.

I don’t understand this sub-setting. I would subset the universe for each ratio separately.

EXPRATIO = sum(GP)/sum (GQ_SIZE_EXP_PERS_CNT) where unres = ‘0’ and FOCS_ER_CB_CODE ‘ ‘ and flagA in (‘ ‘, ‘R’)

MAXRATIO = sum(GP)/sum (GQ_SIZE_MAX_PERS_CNT) where unres = ‘0’ and FOCS_ER_CB_CODE = ‘ ‘ and flagB in (‘ ‘,’R’)

Etc. It will be easier to code this way and it will make maximum use of the reported data.

I think you only need three sets of ratios for each of the four variables, so only 12 applicable factors for each GQTYPCUR. I think the conditions on which variables are populated only matter for the business rules at the end.

Commented [JEZ(F2R1]: I added a table at the end of the document to show what I think we should do, how we could spec out the 12 ratios.

i. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
ii. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
iii. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
iv. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
v. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
vi. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

c. We will create 3 ratios for each variable, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):
   i. Sum the GP and GQCURRSIZE value **for the nation.**
   ii. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
   iii. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
   iv. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
   v. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
   vi. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

d. We will create 3 ratios for each variable, one for the national value (**CURRMAXRATIO)**, one for the GQTYPCUR combination (**CURRMAXRATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):
   i. Sum the GP and GQCURRMAXPOP value **for the nation.**
   ii. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
   iii. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
   iv. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
   v. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
   vi. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

B. Assign Good Person Percentile counts.
   a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP)**, one for the GQTYPCUR combination (**MEDGP_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):
      i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**
      ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**
      iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**
         1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**

2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

a. Define MAXPOP variable.

```
     if gqcurrmaxpop > 0 then maxpop = log(gqcurrmaxpop);
     if gqcurrmaxpop = 0 then maxpop = .;
```

b. Define the fitting universe (ratiofile) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 0 and FOCS_ER_CB_CODE = ‘’

c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.

d. Fit and score this model:

```
proc genmod data = ratiofile;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **poisson_count.**

D. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

E. Fold in CES 501 results

F. Fold in CES 301 results

Section 3: Applying Business Rules

The next section assigns the imputed values. It is broken into three sections based on the auxiliary data.

A. GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

GQTYPCUR	Condition (s)	Method	Flag

Commented [JEZ(F3]: You could do a table like this and write instructions that say, do the imputation by GQTYPCUR. If a MAFID meets the set of conditions, use the method to impute the value and set the flag, if not, move to the next row, etc.

104	GQCURRSIZE > 0 and CURRSIZE_RATIO_GQ_ST > 0	CEIL (CURRSIZE_RATIO_GQ_ST * GQCURRSIZE)	GQIMPPATH = 109
104	GQCURRSIZE > 0 and GQCURRSIZE_GQ > 0	CEIL (CURRSIZE_RATIO_GQ * GQCURRSIZE)	GQIMPPATH = 108

B. at least one of GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT is greater than 0, but they all are not (since it is covered in the case above

C. GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are all 0.

RATIOS:

Create the following ratios by summing values for all IDs where unres = 0 and FOCS_ER_CB_CODE = ‘ ‘. Use the table to determine the level for the ratio and any additional conditions. For example,

$$EXPRATIO_{GQ\_ST} = \frac{\sum_i GP}{\sum_i GQ\_SIZE\_EXP\_PERS\_CNT}$$

$$where\ i\ in\ GQTYPCUR\ and\ BCUSTATEFP\ and\ FLAGA\ in\ ('\ \ ','\ R')$$

Ratio	Numerator	Denominator	Level	Condition
EXPRATIO_GQ_ST	SUM(GP)	SUM(GQ_SIZE_EXP_PERS_CNT)	GQTYPCUR*BCUSTATEFP	FLAGA in (‘ ‘,’R’)
EXPRATIO_GQ	SUM(GP)	SUM(GQ_SIZE_EXP_PERS_CNT)	GQTYPCUR	FLAGA in (‘ ‘,’R’)
EXPRATIO	SUM(GP)	SUM(GQ_SIZE_EXP_PERS_CNT)	All MAFIDs meeting conditions	FLAGA in (‘ ‘, ‘R’)
MAXRATIO_GQ_ST	SUM(GP)	SUM(GQ_SIZE_MAX_PERS_CNT)	GQTYPCUR*BCUSTATEFP	FLAGB in (‘ ‘,’R’)
…				

Andrew Keller, Julianne Zamora, Tim Kennel

December 23, 2020

2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input Files:

1. /sampling/eb/kelle321/gq_mafid_cnts_121920_geo_cdl.sas7bdat
2. /sampling/share/hbparm.sas7bdat
3. CES 501 results
4. CES 301 results

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file, referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.

Rename GQ_INITIAL_STATUS to GQ_PRE_STATUS.
Rename GQ_INITIAL_UNRES to GQ_PRE_UNRES.
Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file, referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.
D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
    a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

Section 2: HB Edits

A. Calculate Ratios for editing.

a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘), then

i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT

ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT

iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE

iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP

b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.

a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR

b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100

7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

   a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

   b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

   c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

   d. For each MAFID, transform the ratio to create **SVALUE**.

      i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

      ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)

   e. For each MAFID, transform SVALUE to create **EVALUE**.

      i. EVALUE = SVALUE * max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}$^{0.5}$

      ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

   f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

      i. **E_Q1** = first quartile EVALUE

      ii. **E_MED** = median EVALUE

      iii. **E_Q3** = third quartile EVALUE

   g. For each GQTYPE, define upper and lower bounds.

      i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

      ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

      iii. **LOWER_C1** = E_MED – C1 * D_Q1

      iv. **LOWER_C2** = E_MED – C2 * D_Q1

      v. **LOWER_C3** = E_MED – C3 * D_Q1

      vi. **UPPER_C1** = E_MED + C1 * D_Q3

      vii. **UPPER_C2** = E_MED + C2 * D_Q3

      viii. **UPPER_C3** = E_MED + C3 * D_Q3

   h. For each MAFID, create **FLAG[X]**.

      i. If EVALUE is missing, FLAG[X] = ‘M’

      ii. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = ‘R’

      iii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = ‘S’

      iv. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = ‘I’

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

   a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (‘ ‘,’O’,’R’) and GQ_INITIAL_UNRES = 0. Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:
   i. If FLAG[X] = ‘I’ and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = ‘S’.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.
   a. If (FLAGA = ‘I’ or FLAGB = ‘I’ or FLAGC = ‘I’ or FLAGD = ‘I’) and IMPUTE_NEEDED ne ‘N’ then
      i. **GP = .**
      ii. **UNRES** = 1
   b. Otherwise,
      i. **GP =** GQ_INITIAL_POP
      ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values
   a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values
      i. Calculate Ratios.
         We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’):
         1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
         2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
         3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
         4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
         5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
         6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
      ii. Assign values. For each MAFID, calculate the following values:
         1. **IMP_RAT_EXP** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO)
         2. **IMP_RAT_EXP_GQ** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ)
         3. **IMP_RAT_EXP_GQ_ST** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values
   i. Calculate Ratios.
      We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):
      1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
      2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
      3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
      4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
      5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
      6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
   ii. Assign values. For each MAFID, calculate the following values:
      1. **IMP_RAT_MAX** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO)
      2. **IMP_RAT_MAX_GQ** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ)
      3. **IMPRAT_MAX_GQ_ST** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values
   i. Calculate Ratios.
      We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):
      1. Sum the GP and GQCURRSIZE value **for the nation.**
      2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
      3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
      4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
      5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
      6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
   ii. Assign values. For each MAFID, calculate the following values:
      1. **IMP_RAT_CURR** = CEIL (GQCURRSIZE*CURRSIZERATIO)
      2. **IMP_RAT_CURR_GQ** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ)
      3. **IMP_RAT_CURR_GQ_ST** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values
   i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (GQCURRMAXPOP*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

a. Define MAXPOP variable.

i. if GQCURRMAXPOP > 0 then **MAXPOP** = log(GQCURRMAXPOP);

ii. if GQCURRMAXPOP = 0 then **MAXPOP** = .;

b. Define the fitting universe (ratiofile) as this: FLAGA in (' ','R') and FLAGB in (' ','R') and FLAGC in (' ','R') and FLAGD in (' ','R') and unres = 0 and FOCS_ER_CB_CODE = ''

c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.

d. Fit and score this model:

```
proc genmod data = ratiofile;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **IMP_POISSON_COUNT.**

Commented [JEZ(F1]: Remove?

D. Fold in CES 501 results

Commented [JEZ(F2]: Residual Method

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table as follows, if IMP_POISSON_COUNT is not missing, assign IMP_GP = IMP_POISSON_COUNT and assign IMP_FLAG = 201. If IMP_POISSON_COUNT is missing, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. Continue on through the table until all MAFIDs in unres = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	**IMP_FLAG**
IMP_POISSON_COUNT	201
IMP_RAT_EXP_GQ_ST	101
IMP_RAT_EXP_GQ	102
IMP_RAT_EXP	103
IMP_RAT_MAX_GQ_ST	104
IMP_RAT_MAX_GQ	105
IMP_RAT_MAX	106
IMP_RAT_CURR_GQ_ST	107
IMP_RAT_CURR_GQ	108
'IMP_RAT_CURR	109
IMP_RAT_CURRMAX_GQ_ST	110
IMP_RAT_CURRMAX_GQ	111
IMP_RAT_CURRMAX	112
MEDGP_GQ_ST	401
MEDGP_GQ	402
MEDGP	403

Commented [JEZ(F3]: Remove?

Section 6: Create Output File

Output GQ_MAFID, adding the following variables:

FLAGA	FLAGB	
FLAGC	FLAGD	
GP	UNRES	
EXPRATIO	EXPRATIO_GQ	EXPRATIO_GQ_ST
IMP_RAT_EXP	IMP_RAT_EXP_GQ	IMP_RAT_EXP_GQ_ST
MAXRATIO	MAXRATIO_GQ	MAXRATIO_GQ_ST
IMP_RAT_MAX	IMP_RAT_MAX_GQ	IMP_RAT_MAX_GQ_ST
CURRRATIO	CURRRATIO_GQ	CURRATIO_GQ_ST
IMP_RAT_CURR	IMP_RAT_CURR_GQ	IMP_RAT_CURR_GQ_ST
MAXCURRRATIO	MAXCURRRATIO_GQ	MAXCURRRATIO_GQ_ST
IMP_RAT_MAXCURR	IMP_RAT_MAXCURR_GQ	IMP_RAT_MAXCURR_GQ_ST
MEDGP	MEDGP_GQ	MEDGP_GQ_ST
IMP_GP	IMP_FLAG	

Name this file gq_mafid_dssd_out.sas7bdat

Andrew Keller, Julianne Zamora, Tim Kennel
December 23, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input Files:

1. /sampling/eb/kelle321/gq_mafid_cnts_121920_geo_cdl.sas7bdat
2. /sampling/share/hbparm.sas7bdat
3. CES 501 results
4. CES 301 results

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file, referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.

Section 2: HB Edits

A. Calculate Ratios for editing.
   a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘) AND GQ_INITIAL_POP > 0 then
      i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT
      ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT
      iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE
      iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP
   b. Otherwise, RATIO[X] should be set to missing.
B. Create HB Parameters.

a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR

b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100
7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

d. For each MAFID, transform the ratio to create **SVALUE**.

i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)

e. For each MAFID, transform SVALUE to create **EVALUE**.

i. EVALUE = SVALUE * max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}$^{0.5}$

ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

i. **E_Q1** = first quartile EVALUE

ii. **E_MED** = median EVALUE

iii. **E_Q3** = third quartile EVALUE

g. For each GQTYPE, define upper and lower bounds.

i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

iii. **LOWER_C1** = E_MED – C1 * D_Q1

iv. **LOWER_C2** = E_MED – C2 * D_Q1

v. **LOWER_C3** = E_MED – C3 * D_Q1

vi. **UPPER_C1** = E_MED + C1 * D_Q3

vii. **UPPER_C2** = E_MED + C2 * D_Q3

viii. **UPPER_C3** = E_MED + C3 * D_Q3

h. For each MAFID, create **FLAG[X]**.

i. If EVALUE is missing, FLAG[X] = 'M'

ii. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = 'R'

iii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'

iv. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0 AND FLAGA not in ('S','I') and FLAGB not in ('S','I') and FLAGC not in ('S','I') and FLAGD not in ('S','I'). Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

i. If FLAG[X] = 'I' and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = 'S'.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

a. If (FLAGA = 'I' or FLAGB = 'I' or FLAGC = 'I' or FLAGD = 'I') and IMPUTE_NEEDED ne 'N' then
    i. **GP = .**
    ii. **UNRES** = 1
b. Otherwise,
    i. **GP =** GQ_INITIAL_POP
    ii. **UNRES** = GQ_INITIAL_UNRES

<u>Section 4: Create Imputed Values</u>

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values
    a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values
        i. Calculate Ratios.
        We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):
            1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
            2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
            3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
            4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
            5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
            6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
        ii. Assign values. For each MAFID, calculate the following values:
            1. **IMP_RAT_EXP** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO)
            2. **IMP_RAT_EXP_GQ** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ)
            3. **IMP_RAT_EXP_GQ_ST** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ_ST)

    b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values
        i. Calculate Ratios.
        We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):
            1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**

2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_MAX** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO)
2. **IMP_RAT_MAX_GQ** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ)
3. **IMPRAT_MAX_GQ_ST** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):

1. Sum the GP and GQCURRSIZE value **for the nation.**
2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURR** = CEIL (GQCURRSIZE*CURRSIZERATIO)
2. **IMP_RAT_CURR_GQ** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ)
3. **IMP_RAT_CURR_GQ_ST** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**

        4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
        5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
        6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
    - ii. Assign values. For each MAFID, calculate the following values:
        1. **IMP_RAT_CURRMAX** = CEIL (GQCURRMAXPOP*CURRMAXRATIO)
        2. **IMP_RAT_CURRMAX_GQ** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ)
        3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

- a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’,’R’):
    - i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**
    - ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**
    - iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**
        1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
        2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
        3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.

- a. Define MAXPOP variable.
    - i. if GQCURRMAXPOP > 0 then **MAXPOP** = log(GQCURRMAXPOP);
    - ii. if GQCURRMAXPOP = 0 then **MAXPOP** = .;
- b. Define the fitting universe (ratiofile) as this: FLAGA in (‘ ‘,’R’) and FLAGB in (‘ ‘,’R’) and FLAGC in (‘ ‘,’R’) and FLAGD in (‘ ‘,’R’) and unres = 0 and FOCS_ER_CB_CODE = ‘’
- c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.
- d. Fit and score this model:

```
proc genmod data = ratiofile;
    class gqtypcur;
    model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
```

```
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **IMP_POISSON_COUNT.**

Commented [JEZ(F1]: Remove?

D. Fold in CES 501 results

Commented [JEZ(F2]: Residual Method

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table as follows, if IMP_POISSON_COUNT is not missing, assign IMP_GP = IMP_POISSON_COUNT and assign IMP_FLAG = 201. If IMP_POISSON_COUNT is missing, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. Continue on through the table until all MAFIDs in unres = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	IMP_FLAG
IMP_POISSON_COUNT	201
IMP_RAT_EXP_GQ_ST	101
IMP_RAT_EXP_GQ	102
IMP_RAT_EXP	103
IMP_RAT_MAX_GQ_ST	104
IMP_RAT_MAX_GQ	105
IMP_RAT_MAX	106
IMP_RAT_CURR_GQ_ST	107
IMP_RAT_CURR_GQ	108
IMP_RAT_CURR	109
IMP_RAT_MAXCURR_GQ_ST	110
IMP_RAT_MAXCURR_GQ	111
IMP_RAT_MAXCURR	112
MEDGP_GQ_ST	401
MEDGP_GQ	402
MEDGP	403

Commented [JEZ(F3]: Remove?

Section 6: Create Output File

Output GQ_MAFID, adding the following variables, renaming GP to GP_HB:

FLAGA	FLAGB	
FLAGC	FLAGD	
GP_HB	UNRES	
EXPRATIO	EXPRATIO_GQ	EXPRATIO_GQ_ST
IMP_RAT_EXP	IMP_RAT_EXP_GQ	IMP_RAT_EXP_GQ_ST

MAXRATIO	MAXRATIO_GQ	MAXRATIO_GQ_ST
IMP_RAT_MAX	IMP_RAT_MAX_GQ	IMP_RAT_MAX_GQ_ST
CURRRATIO	CURRRATIO_GQ	CURRATIO_GQ_ST
IMP_RAT_CURR	IMP_RAT_CURR_GQ	IMP_RAT_CURR_GQ_ST
MAXCURRRATIO	MAXCURRRATIO_GQ	MAXCURRRATIO_GQ_ST
IMP_RAT_MAXCURR	IMP_RAT_MAXCURR_GQ	IMP_RAT_MAXCURR_GQ_ST
MEDGP	MEDGP_GQ	MEDGP_GQ_ST
IMP_GP	IMP_FLAG	

Name this file gq_mafid_dssd_out.sas7bdat

Andrew Keller, Julianne Zamora, Tim Kennel
December 23, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into three sections:

1. Defining the Unresolved Cases Eligible for GQ Size Imputation
2. Developing the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type
3. Assign Business Rules to choose between the imputation methods to assign a final imputed value

Input Files:

1. /sampling/eb/kelle321/gq_mafid_cnts_121920_geo_cdl.sas7bdat
2. /sampling/share/hbparm.sas7bdat
3. CES 501 results
4. CES 301 results

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file, referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.

Rename GQ_INITIAL_STATUS to GQ_PRE_STATUS.
Rename GQ_INITIAL_UNRES to GQ_PRE_UNRES.
Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file, referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.
D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
   a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

<u>Section 2: HB Edits</u>

A. Calculate Ratios for editing.

   a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘), then

      i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT

      ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT

      iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE

      iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP

   b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.

   a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR

   b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100

7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

d. For each MAFID, transform the ratio to create **SVALUE**.

i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)

e. For each MAFID, transform SVALUE to create **EVALUE**.

i. EVALUE = SVALUE * max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}$^{0.5}$

ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

i. **E_Q1** = first quartile EVALUE

ii. **E_MED** = median EVALUE

iii. **E_Q3** = third quartile EVALUE

g. For each GQTYPE, define upper and lower bounds.

i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

iii. **LOWER_C1** = E_MED – C1 * D_Q1

iv. **LOWER_C2** = E_MED – C2 * D_Q1

v. **LOWER_C3** = E_MED – C3 * D_Q1

vi. **UPPER_C1** = E_MED + C1 * D_Q3

vii. **UPPER_C2** = E_MED + C2 * D_Q3

viii. **UPPER_C3** = E_MED + C3 * D_Q3

h. For each MAFID, create **FLAG[X]**.

i. If EVALUE is missing, FLAG[X] = ‘M’

ii. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = ‘R’

iii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = ‘S’

iv. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = ‘I’

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (‘ ‘,’O’,’R’) and GQ_INITIAL_UNRES = 0. Assign these values as **GP_10** and **GP_90** respectively.

   b. For each MAFID and FLAG[X] make the following update:
      i. If FLAG[X] = 'I' and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = 'S'.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.
   a. If (FLAGA = 'I' or FLAGB = 'I' or FLAGC = 'I' or FLAGD = 'I') and IMPUTE_NEEDED ne 'N' then
      i. **GP = .**
      ii. **UNRES** = 1
   b. Otherwise,
      i. **GP =** GQ_INITIAL_POP
      ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values
   a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values
      i. Calculate Ratios.
         We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):
         1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
         2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
         3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
         4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
         5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
         6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
      ii. Assign values. For each MAFID, calculate the following values:
         1. **IMP_RAT_EXP** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO)
         2. **IMP_RAT_EXP_GQ** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ)
         3. **IMP_RAT_EXP_GQ_ST** = CEIL (GQ_SIZE_EXP_PERS_CNT*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values
    i. Calculate Ratios.
    We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):
        1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
        2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
        3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
        4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
        5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
        6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
    ii. Assign values. For each MAFID, calculate the following values:
        1. **IMP_RAT_MAX** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO)
        2. **IMP_RAT_MAX_GQ** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ)
        3. **IMPRAT_MAX_GQ_ST** = CEIL (GQ_SIZE_MAX_PERS_CNT*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values
    i. Calculate Ratios.
    We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):
        1. Sum the GP and GQCURRSIZE value **for the nation.**
        2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
        3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
        4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
        5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
        6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
    ii. Assign values. For each MAFID, calculate the following values:
        1. **IMP_RAT_CURR** = CEIL (GQCURRSIZE*CURRSIZERATIO)
        2. **IMP_RAT_CURR_GQ** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ)
        3. **IMP_RAT_CURR_GQ_ST** = CEIL (GQCURRSIZE*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values
    i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (GQCURRMAXPOP*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (GQCURRMAXPOP*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are <u>**all**</u> greater than 0.

a. Define MAXPOP variable.

i. if GQCURRMAXPOP > 0 then **MAXPOP** = log(GQCURRMAXPOP);

ii. if GQCURRMAXPOP = 0 then **MAXPOP** = .;

b. Define the fitting universe (ratiofile) as this: FLAGA in (‘ ‘,’R’) and FLAGB in (‘ ‘,’R’) and FLAGC in (‘ ‘,’R’) and FLAGD in (‘ ‘,’R’) and unres = 0 and FOCS_ER_CB_CODE = ‘’

c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.

d. Fit and score this model:

```
proc genmod data = ratiofile;
     class gqtypcur;
     model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT
GQ_SIZE_EXP_PERS_CNT /
           link = log d = poisson offset = maxpop maxiter = 500;
  store params;
     output out = poi_pred PREDICTED = pr_size;
run;

proc plm source=params;
  score data = nomaxscore out=nomaxscoreout/ ilink;
run;
```

e. Take the ceiling function of the predicted count. Call this **IMP_POISSON_COUNT.**

Commented [JEZ(F1]: Remove?

D. Fold in CES 501 results

Commented [JEZ(F2]: Residual Method

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table as follows, if IMP_POISSON_COUNT is not missing, assign IMP_GP = IMP_POISSON_COUNT and assign IMP_FLAG = 201. If IMP_POISSON_COUNT is missing, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. Continue on through the table until all MAFIDs in unres = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	**IMP_FLAG**
IMP_POISSON_COUNT	201
IMP_RAT_EXP_GQ_ST	101
IMP_RAT_EXP_GQ	102
IMP_RAT_EXP	103
IMP_RAT_MAX_GQ_ST	104
IMP_RAT_MAX_GQ	105
IMP_RAT_MAX	106
IMP_RAT_CURR_GQ_ST	107
IMP_RAT_CURR_GQ	108
‘IMP_RAT_CURR	109
IMP_RAT_CURRMAX_GQ_ST	110
IMP_RAT_CURRMAX_GQ	111
IMP_RAT_CURRMAX	112
MEDGP_GQ_ST	401
MEDGP_GQ	402
MEDGP	403

Commented [JEZ(F3]: Remove?

Section 6: Create Output File

Output GQ_MAFID, adding the following variables:

FLAGA	FLAGB	
FLAGC	FLAGD	
GP	UNRES	
EXPRATIO	EXPRATIO_GQ	EXPRATIO_GQ_ST
IMP_RAT_EXP	IMP_RAT_EXP_GQ	IMP_RAT_EXP_GQ_ST
MAXRATIO	MAXRATIO_GQ	MAXRATIO_GQ_ST
IMP_RAT_MAX	IMP_RAT_MAX_GQ	IMP_RAT_MAX_GQ_ST
CURRRATIO	CURRRATIO_GQ	CURRATIO_GQ_ST
IMP_RAT_CURR	IMP_RAT_CURR_GQ	IMP_RAT_CURR_GQ_ST
MAXCURRRATIO	MAXCURRRATIO_GQ	MAXCURRRATIO_GQ_ST
IMP_RAT_MAXCURR	IMP_RAT_MAXCURR_GQ	IMP_RAT_MAXCURR_GQ_ST
MEDGP	MEDGP_GQ	MEDGP_GQ_ST
IMP_GP	IMP_FLAG	

Name this file gq_mafid_dssd_out.sas7bdat

Andrew Keller, Julianne Zamora, Tim Kennel
December ~~23~~24, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into six sections:

1. Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation
2. Running HB Edits
3. Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation
4. Creating Imputed Values
5. Apply Ordering to Select Final Imputed Value
6. Create Output File

Input Files:

1. /sampling/eb/kelle321/gq_mafid_cnts_121920_geo_cdl2.sas7bdat
2. /sampling/share/hbparm.sas7bdat
~~2.~~3. /sampling/share/gqmafid_undup_12220_more.sas7bdat
~~3.~~4. CES 501 results
~~4. CES 301 results~~

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file (gq_mafid_cnts_121920_geo_cdl2.sas7bdat), referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.

Rename GQ_INITIAL_STATUS to GQ_PRE_STATUS.
Rename GQ_INITIAL_UNRES to GQ_PRE_UNRES.
Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file (gqmafid_undup_12220_more.sas7bdat), referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.

D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP

  a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

Section 2: HB Edits

A. Calculate Ratios for editing.

  a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘) and GQ_INITIAL_POP > 0, then

    i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT

    ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT

    iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE

    iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP

  b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.

  a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR

  b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100

3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100
7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

- a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.
- b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.
- c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.
- d. For each MAFID, transform the ratio to create **SVALUE**.
    - i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])
    - ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)
- e. For each MAFID, transform SVALUE to create **EVALUE**.
    - i. EVALUE = SVALUE * max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}$^{0.5}$
    - ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.
- f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.
    - i. **E_Q1** = first quartile EVALUE
    - ii. **E_MED** = median EVALUE
    - iii. **E_Q3** = third quartile EVALUE
- g. For each GQTYPE, define upper and lower bounds.
    - i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}
    - ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}
    - iii. **LOWER_C1** = E_MED – C1 * D_Q1
    - iv. **LOWER_C2** = E_MED – C2 * D_Q1
    - v. **LOWER_C3** = E_MED – C3 * D_Q1
    - vi. **UPPER_C1** = E_MED + C1 * D_Q3
    - vii. **UPPER_C2** = E_MED + C2 * D_Q3
    - viii. **UPPER_C3** = E_MED + C3 * D_Q3
- h. For each MAFID, create **FLAG[X]**.
    - i. If EVALUE is missing, FLAG[X] = 'M'
    - ii. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = 'R'
    - iii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'
    - iv. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0 and FLAGA not in ('S','I') and FLAGB not in ('S','I') and FLAGC not in ('S','I') and FLAGD not in ('S','I'). Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

   i. If FLAG[X] = 'I' and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = 'S'.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

   a. If (FLAGA = 'I' or FLAGB = 'I' or FLAGC = 'I' or FLAGD = 'I') and IMPUTE_NEEDED ne 'N' then

      i. **GP = .**

      ii. **UNRES** = 1

   b. Otherwise,

      i. **GP =** GQ_INITIAL_POP

      ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

   a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values

      i. Calculate Ratios.

      We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):

         1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
         2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
         3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
         4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
         5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
         6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

      ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_EXP_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGA in (' ','R'). Assign these values as **EXP_PERS_10** and **EXP_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_EXP_PERS_CNT.

1. Assign **EXP_PERS_TRUNC** = GQ_SIZE_EXP_PERS_CNT
2. If GQ_SIZE_EXP_PERS_CNT > EXP_PERS_90 then set **EXP_PERS_TRUNC** = EXP_PERS_90

~~6.~~3. If GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_EXP_PERS_CNT < EXP_PERS_10 then set **EXP_PERS_TRUNC** = EXP_PERS_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_EXP** = CEIL (~~GQ_SIZE_EXP_PERS_CNT~~EXP_PERS_TRUNC*EXPRATIO)
2. **IMP_RAT_EXP_GQ** = CEIL (~~GQ_SIZE_EXP_PERS_CNT~~EXP_PERS_TRUNC*EXPRATIO_GQ)
3. **IMP_RAT_EXP_GQ_ST** = CEIL (EXP_PERS_TRUNC~~GQ_SIZE_EXP_PERS_CNT~~*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):

1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_MAX_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGB in (' ','R'). Assign these values as **MAX_PERS_10** and **MAX_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_MAX_PERS_CNT.

1. Assign **MAX_PERS_TRUNC** = GQ_SIZE_MAX_PERS_CNT
2. If GQ_SIZE_MAX_PERS_CNT > MAX_PERS_90 then set **MAX_PERS_TRUNC** = MAX_PERS_90

~~7.~~3. If GQ_SIZE_MAX_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT < MAX_PERS_10 then set **MAX_PERS_TRUNC** = MAX_PERS_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_MAX** = CEIL (~~GQ_SIZE_MAX_PERS_CNT~~MAX_PERS_TRUNC*MAXRATIO)
2. **IMP_RAT_MAX_GQ** = CEIL (~~GQ_SIZE_MAX_PERS_CNT~~MAX_PERS_TRUNC*MAXRATIO_GQ)
3. **IMPRAT_MAX_GQ_ST** = CEIL (~~GQ_SIZE_MAX_PERS_CNT~~MAX_PERS_TRUNC*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):

1. Sum the GP and GQCURRSIZE value **for the nation.**
2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRSIZE for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGC in (' ','R'). Assign these values as **CURRSIZE_10** and **CURRSIZE_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRSIZE.

1. Assign **CURRSIZE_TRUNC** = GQCURRSIZE
2. If GQCURRSIZE > CURRSIZE_90 then set **CURRSIZE_TRUNC** = CURRSIZE_90
~~6.~~3. If GQCURRSIZE > 0 and GQCURRSIZE < CURRSIZE_10 then set **CURRSIZE_TRUNC** = CURRSIZE_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURR** = CEIL (CURRSIZE_TRUNC~~GQCURRSIZE~~*CURRSIZERATIO)
2. **IMP_RAT_CURR_GQ** = CEIL (CURRSIZE_TRUNC~~GQCURRSIZE~~*CURRSIZERATIO_GQ)
3. **IMP_RAT_CURR_GQ_ST** = CEIL (CURRSIZE_TRUNC~~GQCURRSIZE~~*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagD in (‘’,’R’):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRMAXPOP~~SIZE~~ for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGD in (‘ ‘,’R’). Assign these values as **CURRMAX_10** and **CURRMAX_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRMAXPOP.

1. Assign **CURRMAX_TRUNC** = GQCURRMAXPOP
2. If GQCURRMAXPOP > CURRMAX_90 then set **CURRMAX_TRUNC** = CURRMAX_90
3. If GQCURRMAXPOP > 0 and GQCURRMAXPOP < CURRMAX_10 then set **CURRMAX_TRUNC** = CURRMAX_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (~~GQCURRMAXPOP~~CURRMAX_TRUNC*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (CURRMAX_TRUNCT~~GQCURRMAXPOP~~*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (CURRMAX_TRUNC~~GQCURRMAXPOP~~*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’,’R’):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**

2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

iv. Assign values. For each MAFID, calculate the following values:

1. **IMP_MEDGP_GQ_ST** = CEIL(MEDGP_GQ_ST)
2. **IMP_MEDGP_GQ** = CEIL(MEDGP_GQ)
3. **IMP_MEDGP** = CEIL(MEDGP)

~~C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.~~

~~a. Define MAXPOP variable.~~

~~i. if GQCURRMAXPOP > 0 then **MAXPOP** = log(GQCURRMAXPOP);~~

~~ii. if GQCURRMAXPOP = 0 then **MAXPOP** = .;~~

~~b. Define the fitting universe (ratiofile) as this: FLAGA in (' ','R') and FLAGB in (' ','R') and FLAGC in (' ','R') and FLAGD in (' ','R') and unres = 0 and FOCS_ER_CB_CODE = ''~~

~~c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.~~

~~d. Fit and score this model:~~

```
~~proc genmod data = ratiofile;~~
~~      class gqtypcur;~~
~~      model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT~~
~~GQ_SIZE_EXP_PERS_CNT /~~
~~            link = log d = poisson offset = maxpop maxiter = 500;~~
~~  store params;~~
~~      output out = poi_pred PREDICTED = pr_size;~~
~~run;~~

~~proc plm source=params;~~
~~  score data = nomaxscore out=nomaxscoreout/ ilink;~~
~~run;~~
```

~~e. Take the ceiling function of the predicted count. Call this **IMP_POISSON_COUNT.**~~

~~D.~~C. Fold in CES 501 results

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table hierarchically as follows, if IMP_POISSON_COUNT is not missing, assign IMP_GP = IMP_POISSON_COUNT and assign IMP_FLAG = 201. If IMP_POISSON_COUNT is missing, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. Continue on through the table until all MAFIDs with a ~~in~~ unres = 1 have a value for IMP_GP and IMP_FLAG.

Formatted: Font: Not Bold

Commented [JEZ(F1]: Remove?

Commented [JEZ(F2]: Residual Method

IMP_GP	IMP_FLAG
~~IMP_POISSON_COUNT~~	~~201~~
IMP_RAT_EXP_GQ_ST	101
IMP_RAT_EXP_GQ	102
IMP_RAT_EXP	103
IMP_RAT_MAX_GQ_ST	104
IMP_RAT_MAX_GQ	105
IMP_RAT_MAX	106
IMP_RAT_CURR_GQ_ST	107
IMP_RAT_CURR_GQ	108
‘IMP_RAT_CURR	109
IMP_RAT_CURRMAX_GQ_ST	110
IMP_RAT_CURRMAX_GQ	111
IMP_RAT_CURRMAX	112
MEDGP_GQ_ST	401
MEDGP_GQ	402
MEDGP	403

Section 6: Create Output File

Output GQ_MAFID, adding the following variables:

MAFID		
FLAGA	FLAGB	
FLAGC	FLAGD	
GP	UNRES	
EXPRATIO	EXPRATIO_GQ	EXPRATIO_GQ_ST
EXP_PERS_10	EXP_PERS_90	EXP_PERS_TRUNC
IMP_RAT_EXP	IMP_RAT_EXP_GQ	IMP_RAT_EXP_GQ_ST
MAXRATIO	MAXRATIO_GQ	MAXRATIO_GQ_ST
MAX_PERS_10	MAX_PERS_90	MAX_PERS_TRUNC
IMP_RAT_MAX	IMP_RAT_MAX_GQ	IMP_RAT_MAX_GQ_ST
CURRRATIO	CURRRATIO_GQ	CURRATIO_GQ_ST
CURRSIZE_10	CURRSIZE_90	CURRSIZE_TRUNC
IMP_RAT_CURR	IMP_RAT_CURR_GQ	IMP_RAT_CURR_GQ_ST
MAXCURRRATIO	MAXCURRRATIO_GQ	MAXCURRRATIO_GQ_ST
CURRMAX_10	CURRMAX_90	CURRMAX_TRUNC
IMP_RAT_~~MAX~~CURRMAX	IMP_RAT_~~MAX~~CURRMAX_GQ	IMP_RAT_~~MAX~~CURRMAX_GQ_ST
IMP_MEDGP	IMP_MEDGP_GQ	IMP_MEDGP_GQ_ST
IMP_GP	IMP_FLAG	
GQCURRMAXPOP		
GQCURRSIZE		
GQ_SIZE_EXP_PERS_CNT		
GQ_SIZE_MAX_PERS_CNT		

Name this file gq_mafid_dssd_out.sas7bdat

Andrew Keller, Julianne Zamora, Tim Kennel
December ~~23~~24, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into six sections:

1. Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation
2. Running HB Edits
3. Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation
4. Creating Imputed Values
5. Apply Ordering to Select Final Imputed Value
6. Create Output File

Input Files:

1. /sampling/eb/kelle321/gq_mafid_cnts_121920_geo_cdl2.sas7bdat
2. /sampling/share/hbparm.sas7bdat
~~2.~~3. /sampling/share/gqmafid_undup_12220_more.sas7bdat
~~3.~~4. CES 501 results
~~4. CES 301 results~~

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file (gq_mafid_cnts_121920_geo_cdl2.sas7bdat), referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.

Rename GQ_INITIAL_STATUS to GQ_PRE_STATUS.
Rename GQ_INITIAL_UNRES to GQ_PRE_UNRES.
Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file (gqmafid_undup_12220_more.sas7bdat), referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.

D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
   a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

Section 2: HB Edits

A. Calculate Ratios for editing.
   a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in ('O','R',' ') and GQ_INITIAL_POP > 0, then
      i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT
      ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT
      iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE
      iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP
   b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.
   a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR
   b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100

3	D	75	100	175
4	D	25	50	100
5	D	100	125	175
6	D	25	50	100
7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

d. For each MAFID, transform the ratio to create **SVALUE**.

i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO)

e. For each MAFID, transform SVALUE to create **EVALUE**.

i. EVALUE = SVALUE * max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}$^{0.5}$

ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

i. **E_Q1** = first quartile EVALUE

ii. **E_MED** = median EVALUE

iii. **E_Q3** = third quartile EVALUE

g. For each GQTYPE, define upper and lower bounds.

i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

iii. **LOWER_C1** = E_MED – C1 * D_Q1

iv. **LOWER_C2** = E_MED – C2 * D_Q1

v. **LOWER_C3** = E_MED – C3 * D_Q1

vi. **UPPER_C1** = E_MED + C1 * D_Q3

vii. **UPPER_C2** = E_MED + C2 * D_Q3

viii. **UPPER_C3** = E_MED + C3 * D_Q3

h. For each MAFID, create **FLAG[X]**.

i. If EVALUE is missing, FLAG[X] = 'M'

ii. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = 'R'

iii. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'

iv. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0 and FLAGA not in ('S','I') and FLAGB not in ('S','I') and FLAGC not in ('S','I') and FLAGD not in ('S','I'). Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

   i. If FLAG[X] = 'I' and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = 'S'.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

a. If (FLAGA = 'I' or FLAGB = 'I' or FLAGC = 'I' or FLAGD = 'I') and IMPUTE_NEEDED ne 'N' then

   i. **GP = .**

   ii. **UNRES** = 1

b. Otherwise,

   i. **GP =** GQ_INITIAL_POP

   ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values

   i. Calculate Ratios.

   We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):

   1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
   2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
   3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
   4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
   5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
   6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

   ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_EXP_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGA in (‘ ‘,’R’). Assign these values as **EXP_PERS_10** and **EXP_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_EXP_PERS_CNT.

1. Assign **EXP_PERS_TRUNC** = GQ_SIZE_EXP_PERS_CNT
2. If GQ_SIZE_EXP_PERS_CNT > EXP_PERS_90 then set **EXP_PERS_TRUNC** = EXP_PERS_90

~~6.~~3. If GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_EXP_PERS_CNT < EXP_PERS_10 then set **EXP_PERS_TRUNC** = EXP_PERS_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_EXP** = CEIL (~~GQ_SIZE_EXP_PERS_CNT~~EXP_PERS_TRUNC*EXPRATIO)
2. **IMP_RAT_EXP_GQ** = CEIL (~~GQ_SIZE_EXP_PERS_CNT~~EXP_PERS_TRUNC*EXPRATIO_GQ)
3. **IMP_RAT_EXP_GQ_ST** = CEIL (EXP_PERS_TRUNC~~GQ_SIZE_EXP_PERS_CNT~~*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagB in (‘’,’R’):

1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_MAX_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGB in (‘ ‘,’R’). Assign these values as **MAX_PERS_10** and **MAX_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_MAX_PERS_CNT.

1. Assign **MAX_PERS_TRUNC** = GQ_SIZE_MAX_PERS_CNT
2. If GQ_SIZE_MAX_PERS_CNT > MAX_PERS_90 then set **MAX_PERS_TRUNC** = MAX_PERS_90

~~7.~~3. If GQ_SIZE_MAX_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT < MAX_PERS_10 then set **MAX_PERS_TRUNC** = MAX_PERS_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_MAX** = CEIL (~~GQ_SIZE_MAX_PERS_CNT~~MAX_PERS_TRUNC*MAXRATIO)
2. **IMP_RAT_MAX_GQ** = CEIL (~~GQ_SIZE_MAX_PERS_CNT~~MAX_PERS_TRUNC*MAXRATIO_GQ)
3. **IMPRAT_MAX_GQ_ST** = CEIL (~~GQ_SIZE_MAX_PERS_CNT~~MAX_PERS_TRUNC*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagC in ('','R'):

1. Sum the GP and GQCURRSIZE value **for the nation.**
2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRSIZE for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGC in (' ','R'). Assign these values as **CURRSIZE_10** and **CURRSIZE_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRSIZE.

1. Assign **CURRSIZE_TRUNC** = GQCURRSIZE
2. If GQCURRSIZE > CURRSIZE_90 then set **CURRSIZE_TRUNC** = CURRSIZE_90

~~6.~~3. If GQCURRSIZE > 0 and GQCURRSIZE < CURRSIZE_10 then set **CURRSIZE_TRUNC** = CURRSIZE_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURR** = CEIL (CURRSIZE_TRUNC~~GQCURRSIZE~~*CURRSIZERATIO)
2. **IMP_RAT_CURR_GQ** = CEIL (CURRSIZE_TRUNC~~GQCURRSIZE~~*CURRSIZERATIO_GQ)
3. **IMP_RAT_CURR_GQ_ST** = CEIL (CURRSIZE_TRUNC~~GQCURRSIZE~~*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRMAXPOP~~SIZE~~ for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGD in (' ','R'). Assign these values as **CURRMAX_10** and **CURRMAX_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRMAXPOP.

1. Assign **CURRMAX_TRUNC** = GQCURRMAXPOP
2. If GQCURRMAXPOP > CURRMAX_90 then set **CURRMAX_TRUNC** = CURRMAX_90
3. If GQCURRMAXPOP > 0 and GQCURRMAXPOP < CURRMAX_10 then set **CURRMAX_TRUNC** = CURRMAX_10.

~~ii.~~iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (~~GQCURRMAXPOP~~CURRMAX_TRUNC*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (CURRMAX_TRUNCT~~GQCURRMAXPOP~~*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (CURRMAX_TRUNC~~GQCURRMAXPOP~~*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**

2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

Formatted: Font: Not Bold

iv. Assign values. For each MAFID, calculate the following values:

1. **IMP_MEDGP_GQ_ST** = CEIL(MEDGP_GQ_ST)
2. **IMP_MEDGP_GQ** = CEIL(MEDGP_GQ)
3. **IMP_MEDGP** = CEIL(MEDGP)

~~C. Run a Poisson regression model to get predicted good person counts on counts for GQ where GQCURRMAXPOP GQCURRSIZE GQ_SIZE_EXP_PERS_CNT GQ_SIZE_MAX_PERS_CNT are **all** greater than 0.~~

~~a. Define MAXPOP variable.~~

~~i. if GQCURRMAXPOP > 0 then **MAXPOP** = log(GQCURRMAXPOP);~~

~~ii. if GQCURRMAXPOP = 0 then **MAXPOP** = .;~~

~~b. Define the fitting universe (ratiofile) as this: FLAGA in (' ','R') and FLAGB in (' ','R') and FLAGC in (' ','R') and FLAGD in (' ','R') and unres = 0 and FOCS_ER_CB_CODE = ''~~

~~c. Define the scoring universe (nomaxscore) as this: GQCURRMAXPOP > 0 and GQCURRSIZE > 0 and GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT > 0 and unres = 1.~~

~~d. Fit and score this model:~~

```
~~proc genmod data = ratiofile;~~
~~      class gqtypcur;~~
~~      model gp = gqtypcur gqcurrsize GQ_SIZE_MAX_PERS_CNT~~
~~GQ_SIZE_EXP_PERS_CNT /~~
~~                  link = log d = poisson offset = maxpop maxiter = 500;~~
~~   store params;~~
~~      output out = poi_pred PREDICTED = pr_size;~~
~~run;~~

~~proc plm source=params;~~
~~   score data = nomaxscore out=nomaxscoreout/ ilink;~~
~~run;~~
```

~~e. Take the ceiling function of the predicted count. Call this **IMP_POISSON_COUNT.**~~

Commented [JEZ(F1]: Remove?

C. Residual method: using a hybrid of the ratio imputes created in the previous step, a percentile method based on Greek/non-Greek status, and allocation of a facility-level residual to individual MAFIDs.

a. Ingest the file referred to as **MAFID_FRAT_SORO**

i. On this file **FLAG_GREEK_LETTER**=1 indicates that GQ has been identified as a fraternity or sorority house. Otherwise **FLAG_GREEK_LETTER**=0.

b. Ingest the file referred to as **UNITID_MAFID_LINKS**.

i. When reading in **UNITID_MAFID_LINKS,** keep only the variables **MAFID, UNITID, MATCH_STEP_NUM**, and **ROOMCAP.**

ii. Note: for records with **MATCH_STEP_NUM**=-1, **UNITID** will be missing.

iii. Note: for records with the same value of UNITID, ROOMCAP will be the same.

c. Merge **MAFID_FRAT_SORO** and **UNITID_MAFID_LINKS** to ***GQ_MAFID***, merging on MAFID, and keeping only records that are in ***GQ_MAFID.***

i. Note: For records that match, this should be a 1-to-1 match (MAFID should be unique in each of the 3 datasets).

ii. Note: only records with GQCURTYP=501 in ***GQ_MAFID*** should match to either of the other 2 datasets.

d. Select the subset of the merged dataset from the previous step with GQCURTYP=501.

i. NOTE: In this spec we will refer to this subset of the data as **GQ_COUNTS_ROOMCAP_GREEK**. This is only an intermediate dataset, which will be merged back to the **GQ_MAFID** dataset at the end of this section of the spec (section 5.D).

e. Using GQ_COUNTS_ROOM_CAP_GREEK and the ratio impute variables created in section 4.A, create a temporary impute variable IMP_GP_TEMP using the hierarchy shown in the following table. If IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP_TEMP= IMP_RAT_EXP_GQ_ST and set ALREADY_IMPUTED=1. If IMP_RAT_EXP_GQ_ST is missing and IMP_RAT_EXP_GQ is not missing, assign IMP_GP_TEMP= IMP_RAT_EXP_GQ and set ALREADY_IMPUTED=1. Continue through the table until all the variables in the table have been exhausted. For any remaining MAFIDs for which a value has not been assigned to IMP_GP_TEMP, set ALREADY_IMPUTED=0;

IMP_GP_TEMP assignment hierarchy
IMP_RAT_EXP_GQ_ST
IMP_RAT_EXP_GQ
IMP_RAT_MAX_GQ_ST
IMP_RAT_MAX_GQ
IMP_RAT_CURR_GQ_ST
IMP_RAT_CURR_GQ
IMP_RAT_CURRMAX_GQ_ST
IMP_RAT_CURRMAX_GQ

f. Using only MAFIDs in **GQ_COUNTS_ROOMCAP_GREEK** with UNRES = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’,’R’), create 3 GP median variables and 3 GP maximum variables:

i. For each UNITID-FLAG_GREEK_LETTER combination with enough MAFIDs:

1. Calculate the median value of GP. Call this **P50_GP_UNIT_BY_GRK**
2. Calculate the maximum value of GP. Call this **MAX_GP_UNIT_BY_GRK.**
3. Merge the P50_GP_UNIT_BY_GRK and MAX_GP_UNIT_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on UNITID and FLAG_GREEK_LETTER.

ii. For each BCUSTATEFP-FLAG_GREEK_LETTER combination with enough MAFIDs:

1. Calculate the median value of GP. Call this **P50_GP_ST_BY_GRK**.
2. Calculate the maximum value of GP. Call this **MAX_GP_ST_BY_GRK**.
3. Merge P50_GP_ST_BY_GRK and MAX_GP_ST_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on BCUSTATEFP-FLAG_GREEK_LETTER combinations.

iii. For each value of FLAG_GREEK_LETTER:

1. Calculate the median value of GP. Call this **P50_GP_BY_GRK.**
2. Calculate the maximum value of GP. Call this **MAX_GP_BY_GRK**.
3. Merge P50_GP_BY_GRK and MAX_BP_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on FLAG_GREEK_LETTER.

g. For MAFIDs for which UNRES=1, FLAG_GREEK_LETTER=1, and ALREADY_IMPUTED=0, assign median Greek imputes to IMP_GP_TEMP and create up to 3 new impute variables using the following hierarchy:

i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:

1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_GRK_UNIT**= IMP_GP_TEMP

ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:

1. assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
2. set ALREADY_IMPUTED=1
3. Assign **MEDGP_GRK_ST**= IMP_GP_TEMP

iii. Otherwise:

1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_GRK**=IMP_GP_TEMP

h. Using **GQ_COUNTS_ROOMCAP_GREEK**, by UNITID, create unit-level sum variables (where a unit corresponds to a single UNITID, which corresponds to a single a university or college)

i. Create unit-level sums (i.e., by UNITID) of GQCURRMAXPOP using only observations where flagD in ('','R'). Note: these are the "good" values of GQCURRMAXPOP. Note that for this sum, we don't care what the value of GP is, even it is a true 0. We are just trying to come up with a maximum number of people that these GQs *could* house, so that we can subtract the sum from the college-level IPEDS ROOMCAP variable. For reference later in the spec, call this sum **UNIT_MAXPOP_SUM**.

ii. Using only the GQs with unres=0 and flagD **not** in ('','R'), by UNITID, create unit-level sums of GP. Call this sum **UNIT_2020POP_SUM**.

iii. Using only the GQs with unres=1 and flagD **not** in ('','R'), by UNITID, create unit-level sums of IMP_GP_TEMP. Call this **UNIT_POP_IMPUTED_SUM**.

iv. Create **UNIT_CAP_SUM** = the unit-level sum of UNIT_MAXPOP_SUM, UNIT_2020POP_SUM, and UNIT_POP_IMPUTED_SUM

i. For each MAFID, calculate UNIT_RESIDUAL = ROOMCAP – UNIT_CAP_SUM (this will be the same value for MAFIDs with the same UNITID)

j. For each MAFID with UNIT_RESIDUAL<=0, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP, and create 3 new (non-Greek) median impute variables using the following hierarchy:

i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:

1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP

ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:

1. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP

iii. Otherwise:

1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

k. For each (non-missing) UNITID with UNIT_RESIDUAL>0, count the MAFIDs associated with that UNITID that have UNRES=1 and ALREADY_IMPUTED=0. Call this count UNIT_RESID_GQ_COUNT.

l. For MAFIDs with UNIT_RESIDUAL>0, UNIT_RESID_GQ_COUNT=1, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy:

i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK, then assign values to IMP_GP_TEMP using the following sub-hierarchy:

1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:

a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK

b. Set ALREADY_IMPUTED=1

c. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP

2. Otherwise (i.e., if P50_GP_UNIT_BY_GRK<=0 or missing), if MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing, then:

a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK

b. Set ALREADY_IMPUTED=1

c. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP

3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:

a. Assign IMP_GP_TEMP= P50_GP_BY_GRK

b. Set ALREADY_IMPUTED=1

c. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL < MAX_GP_UNIT_BY_GRK, then assign values as follows:

1. Assign IMP_GP_TEMP=UNIT_RESIDUAL
2. Set ALREADY_IMPUTED=1
3. Assign **IMP_RESID_1GQ**=IMP_GP_TEMP

m. For MAFIDs with UNIT_RESIDUAL>0, UNIT_RESID_GQ_COUNT>1, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy. (NOTE: steps i.1-i.3 are the same as steps i.1-i.3 in step l above):

i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK, then assign values to IMP_GP_TEMP using the following sub-hierarchy:

1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:

a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK

b. Set ALREADY_IMPUTED=1

c. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP

2. Otherwise (i.e., if P50_GP_UNIT_BY_GRK<=0 or missing), if MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing, then:
    - a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
    - b. Set ALREADY_IMPUTED=1
    - c. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP
3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:
    - a. Assign IMP_GP_TEMP= P50_GP_BY_GRK
    - b. Set ALREADY_IMPUTED=1
    - c. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL < MAX_GP_UNIT_BY_GRK, then assign values as follows:

1. Assign IMP_GP_TEMP=UNIT_RESIDUAL/UNIT_RESID_GQ_COUNT
2. Set ALREADY_IMPUTED=1
3. Assign **IMP_RESID_NGQ**=IMP_GP_TEMP

n. Do a cross-tabulation of the variables UNRES and ALREADY_IMPUTED. If ALREADY_IMPUTED is always 1 when UNRES=1, then imputations have been calculated for all MAFIDS with GQCURTYP 501.

o. Keep the variables **MEDGP_GRK_UNIT, MEDGP_GRK_ST, MEDGP_GRK, MEDGP_nonGRK_UNIT, MEDGP_nonGRK_ST, MEDGP_nonGRK, IMP_RESID_1GQ**, and **IMP_RESID_NGQ.** Drop all other variables created in this section

~~D. Fold in CES 501 results~~

Commented [JEZ(F2]: Residual Method

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table hierarchically as follows, if IMP_POISSON_COUNT is not missing, assign IMP_GP = IMP_POISSON_COUNT and assign IMP_FLAG = 201. If IMP_POISSON_COUNT is missing, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. Continue on through the table until all MAFIDs with a ~~in~~ unres = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	**IMP_FLAG**
~~IMP_POISSON_COUNT~~	~~201~~
IMP_RAT_EXP_GQ_ST	101
IMP_RAT_EXP_GQ	102
IMP_RAT_EXP	103
IMP_RAT_MAX_GQ_ST	104
IMP_RAT_MAX_GQ	105
IMP_RAT_MAX	106
IMP_RAT_CURR_GQ_ST	107
IMP_RAT_CURR_GQ	108
‘IMP_RAT_CURR	109

IMP_RAT_CURRMAX_GQ_ST	110
IMP_RAT_CURRMAX_GQ	111
IMP_RAT_CURRMAX	112
MEDGP_GRK_UNIT	301
MEDGP_GRK_ST	302
MEDGP_GRK	303
MEDGP_nonGRK_UNIT	304
MEDGP_nonGRK_ST	305
MEDGP_nonGRK	306
IMP_RESID_1GQ	307
IMP_RESID_NGQ	308
MEDGP_GQ_ST	401
MEDGP_GQ	402
MEDGP	403

Section 6: Create Output File

Output GQ_MAFID, adding the following variables:

MAFID		
FLAGA	FLAGB	
FLAGC	FLAGD	
GP	UNRES	
EXPRATIO	EXPRATIO_GQ	EXPRATIO_GQ_ST
EXP_PERS_10	EXP_PERS_90	EXP_PERS_TRUNC
IMP_RAT_EXP	IMP_RAT_EXP_GQ	IMP_RAT_EXP_GQ_ST
MAXRATIO	MAXRATIO_GQ	MAXRATIO_GQ_ST
MAX_PERS_10	MAX_PERS_90	MAX_PERS_TRUNC
IMP_RAT_MAX	IMP_RAT_MAX_GQ	IMP_RAT_MAX_GQ_ST
CURRRATIO	CURRRATIO_GQ	CURRATIO_GQ_ST
CURRSIZE_10	CURRSIZE_90	CURRSIZE_TRUNC
IMP_RAT_CURR	IMP_RAT_CURR_GQ	IMP_RAT_CURR_GQ_ST
MAXCURRRATIO	MAXCURRRATIO_GQ	MAXCURRRATIO_GQ_ST
CURRMAX_10	CURRMAX_90	CURRMAX_TRUNC
IMP_RAT_~~MAX~~CURRMAX	IMP_RAT_~~MAX~~CURRMAX_GQ	IMP_RAT_~~MAX~~CURRMAX_GQ_ST
IMP_MEDGP	IMP_MEDGP_GQ	IMP_MEDGP_GQ_ST
IMP_GP	IMP_FLAG	
GQCURRMAXPOP		
GQCURRSIZE		
GQ_SIZE_EXP_PERS_CNT		
GQ_SIZE_MAX_PERS_CNT		

Name this file gq_mafid_dssd_out.sas7bdat

Andrew Keller, Julianne Zamora, Tim Kennel, Kirk White
December 26, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into six sections:

1. Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation
2. Running HB Edits
3. Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation
4. Creating Imputed Values
5. Apply Ordering to Select Final Imputed Value
6. Create Output File

Input Files:

1. /sampling/eb/kelle321/gq_mafid_cnts_121920_geo_cdl2.sas7bdat (GQ_MAFID)
2. /sampling/share/hbparm.sas7bdat (HBPARM)
3. /sampling/share/gqmafid_undup_12220_more.sas7bdat (GQ_DUP_MAFID)
4. /sampling/share/mafid_frat_soro.csv (MAFID_FRAT_SORO)
5. /sampling/share/united_mafid_links.sas7bdat (UNITID_MAFID_LINKS)

Output File: DSSD GQ Imputation File (gq_mafid_dssd_out.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file (gq_mafid_cnts_121920_geo_cdl2.sas7bdat), referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.
D. Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file (gqmafid_undup_12220_more.sas7bdat), referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.
D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
    a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

Section 2: HB Edits

A. Calculate Ratios for editing.

a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘) and GQ_INITIAL_POP > 0, then

i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT

ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT

iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE

iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP

b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.

a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR

b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* (hbparm.sas7bdat) file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175
4	D	25	50	100

5	D	100	125	175
6	D	25	50	100
7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

   a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

   b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

   c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

   d. For each MAFID, transform the ratio to create **SVALUE**.

      i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

      ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO) - 1

   e. For each MAFID, transform SVALUE to create **EVALUE**.

      i. Calculate MAX_INTIAL_POP as max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}

      ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

      iii. EVALUE = SVALUE *(MAX_INITIAL_POP) $^{1/2}$

   f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

      i. **E_Q1** = first quartile EVALUE

      ii. **E_MED** = median EVALUE

      iii. **E_Q3** = third quartile EVALUE

   g. For each GQTYPE, define upper and lower bounds.

      i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

      ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

      iii. **LOWER_C1** = E_MED – C1 * D_Q1

      iv. **LOWER_C2** = E_MED – C2 * D_Q1

      v. **LOWER_C3** = E_MED – C3 * D_Q1

      vi. **UPPER_C1** = E_MED + C1 * D_Q3

      vii. **UPPER_C2** = E_MED + C2 * D_Q3

      viii. **UPPER_C3** = E_MED + C3 * D_Q3

   h. For each MAFID, create **FLAG[X]**.

      i. If EVALUE is missing, FLAG[X] = ‘M’

      ii. Otherwise, apply the following conditions, without nesting (i.e. apply each ‘if’ statement separately).

         1. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = ‘R’

2. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'
3. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0 and FLAGA not in ('S','I') and FLAGB not in ('S','I') and FLAGC not in ('S','I') and FLAGD not in ('S','I') and GQ_INITIAL_POP > 0. Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

i. If FLAG[X] = 'I' and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = 'S'.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

a. If (FLAGA = 'I' or FLAGB = 'I' or FLAGC = 'I' or FLAGD = 'I') and IMPUTE_NEEDED ne 'N' then

i. **GP = .**

ii. **UNRES** = 1

b. Otherwise,

i. **GP =** GQ_INITIAL_POP

ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

a. Calculate GP/GQ_EXP_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQ_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R'):

1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**

4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_EXP_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGA in (' ','R'). Assign these values as **EXP_PERS_10** and **EXP_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_EXP_PERS_CNT.

1. Assign **EXP_PERS_TRUNC** = GQ_SIZE_EXP_PERS_CNT
2. If GQ_SIZE_EXP_PERS_CNT > EXP_PERS_90 then set **EXP_PERS_TRUNC** = EXP_PERS_90
3. If GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_EXP_PERS_CNT < EXP_PERS_10 then set **EXP_PERS_TRUNC** = EXP_PERS_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_EXP** = CEIL (EXP_PERS_TRUNC*EXPRATIO)
2. **IMP_RAT_EXP_GQ** = CEIL (EXP_PERS_TRUNC*EXPRATIO_GQ)
3. **IMP_RAT_EXP_GQ_ST** = CEIL (EXP_PERS_TRUNC*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagB in ('','R'):

1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID
5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_MAX_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGB in (' ','R'). Assign these values as **MAX_PERS_10** and **MAX_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_MAX_PERS_CNT.

1. Assign **MAX_PERS_TRUNC** = GQ_SIZE_MAX_PERS_CNT
2. If GQ_SIZE_MAX_PERS_CNT > MAX_PERS_90 then set **MAX_PERS_TRUNC** = MAX_PERS_90
3. If GQ_SIZE_MAX_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT < MAX_PERS_10 then set **MAX_PERS_TRUNC** = MAX_PERS_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_MAX** = CEIL (MAX_PERS_TRUNC*MAXRATIO)
2. **IMP_RAT_MAX_GQ** = CEIL (MAX_PERS_TRUNC*MAXRATIO_GQ)
3. **IMPRAT_MAX_GQ_ST** = CEIL (MAX_PERS_TRUNC*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagC in (‘’,’R’):

1. Sum the GP and GQCURRSIZE value **for the nation.**
2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRSIZE for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGC in (‘ ‘,’R’). Assign these values as **CURRSIZE_10** and **CURRSIZE_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRSIZE.

1. Assign **CURRSIZE_TRUNC** = GQCURRSIZE
2. If GQCURRSIZE > CURRSIZE_90 then set **CURRSIZE_TRUNC** = CURRSIZE_90
3. If GQCURRSIZE > 0 and GQCURRSIZE < CURRSIZE_10 then set **CURRSIZE_TRUNC** = CURRSIZE_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURR** = CEIL (CURRSIZE_TRUNC*CURRSIZERATIO)
2. **IMP_RAT_CURR_GQ** = CEIL (CURRSIZE_TRUNC*CURRSIZERATIO_GQ)
3. **IMP_RAT_CURR_GQ_ST** = CEIL (CURRSIZE_TRUNC*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagD in ('','R'):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRMAXPOP for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ' ' and FLAGD in (' ','R'). Assign these values as **CURRMAX_10** and **CURRMAX_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRMAXPOP.

1. Assign **CURRMAX_TRUNC** = GQCURRMAXPOP
2. If GQCURRMAXPOP > CURRMAX_90 then set **CURRMAX_TRUNC** = CURRMAX_90
3. If GQCURRMAXPOP > 0 and GQCURRMAXPOP < CURRMAX_10 then set **CURRMAX_TRUNC** = CURRMAX_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (CURRMAX_TRUNC*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (CURRMAX_TRUNC*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (CURRMAX_TRUNC*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = '' and flagA in ('','R') and flagB in ('','R') and flagC in ('','R') and flagD in ('','R'):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**

2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

iv. Assign values. For each MAFID, calculate the following values:

1. **IMP_MEDGP_GQ_ST** = CEIL(MEDGP_GQ_ST)
2. **IMP_MEDGP_GQ** = CEIL(MEDGP_GQ)
3. **IMP_MEDGP** = CEIL(MEDGP)

C. CES method: impute using a hybrid of the ratio imputes created in the previous step, a percentile method based on Greek/non-Greek status, and a facility-level residual allocation method.

a. Ingest the file referred to as **MAFID_FRAT_SORO**

i. On this file **FLAG_GREEK_LETTER**=1 indicates that GQ has been identified as a fraternity or sorority house. Otherwise **FLAG_GREEK_LETTER**=0.

b. Ingest the file referred to as **UNITID_MAFID_LINKS**.

i. When reading in **UNITID_MAFID_LINKS,** keep only the variables **MAFID, UNITID, MATCH_STEP_NUM**, and **ROOMCAP.**

ii. Note: for records with **MATCH_STEP_NUM**=-1, **UNITID** will be missing.

iii. Note: for records with the same value of UNITID, ROOMCAP will be the same.

c. Merge **MAFID_FRAT_SORO** and **UNITID_MAFID_LINKS** to ***GQ_MAFID***, merging on MAFID, and keeping only records that are in ***GQ_MAFID.***

i. Note: For records that match, this should be a 1-to-1 match (MAFID should be unique in each of the 3 datasets).

ii. Note: only records with GQCURTYP=501 in ***GQ_MAFID*** should match to either of the other 2 datasets.

d. Select the subset of the merged dataset from the previous step with GQCURTYP=501.

i. NOTE: In this spec we will refer to this subset of the data as **GQ_COUNTS_ROOMCAP_GREEK**. This is only an intermediate dataset, which will be merged back to the **GQ_MAFID** dataset at the end of this section of the spec (section 5.D).

e. Using GQ_COUNTS_ROOM_CAP_GREEK and the ratio impute variables created in section 4.A, create a temporary impute variable IMP_GP_TEMP using the hierarchy shown in the following table. If IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP_TEMP= IMP_RAT_EXP_GQ_ST and set ALREADY_IMPUTED=1. If IMP_RAT_EXP_GQ_ST is missing and IMP_RAT_EXP_GQ is not missing, assign IMP_GP_TEMP= IMP_RAT_EXP_GQ and set ALREADY_IMPUTED=1. Continue through the table until all the variables in the table have been exhausted. For any remaining MAFIDs for which a value has not been assigned to IMP_GP_TEMP, set ALREADY_IMPUTED=0;

IMP_GP_TEMP assignment hierarchy
IMP_RAT_EXP_GQ_ST
IMP_RAT_EXP_GQ
IMP_RAT_MAX_GQ_ST
IMP_RAT_MAX_GQ

IMP_RAT_CURR_GQ_ST
IMP_RAT_CURR_GQ
IMP_RAT_CURRMAX_GQ_ST
IMP_RAT_CURRMAX_GQ

f. Using only MAFIDs in **GQ_COUNTS_ROOMCAP_GREEK** with UNRES = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’,’R’), create 3 GP median variables and 3 GP maximum variables:

i. For each UNITID-FLAG_GREEK_LETTER combination MAFIDs:

1. Calculate the median value of GP. Call this **P50_GP_UNIT_BY_GRK**
2. Calculate the maximum value of GP. Call this **MAX_GP_UNIT_BY_GRK.**
3. Merge the P50_GP_UNIT_BY_GRK and MAX_GP_UNIT_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on UNITID and FLAG_GREEK_LETTER.
4. Note, these values will be missing if there are not enough observations for the UNITID-FLAG_GREEK_LETTER combination.

~~v.~~ii. For each BCUSTATEFP-FLAG_GREEK_LETTER combination ~~with enough~~ MAFIDs:

1. Calculate the median value of GP. Call this **P50_GP_ST_BY_GRK**.
2. Calculate the maximum value of GP. Call this **MAX_GP_ST_BY_GRK**.
3. Merge P50_GP_ST_BY_GRK and MAX_GP_ST_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on BCUSTATEFP-FLAG_GREEK_LETTER combinations.
4. Note, these values will be missing if there are not enough observations for the BCUSTATEFP-FLAG_GREEK_LETTER combination.

iii. For each value of FLAG_GREEK_LETTER:

1. Calculate the median value of GP. Call this **P50_GP_BY_GRK.**
2. Calculate the maximum value of GP. Call this **MAX_GP_BY_GRK**.
3. Merge P50_GP_BY_GRK and MAX_BP_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on FLAG_GREEK_LETTER.

g. For MAFIDs for which UNRES=1, FLAG_GREEK_LETTER=1, and ALREADY_IMPUTED=0, assign median Greek imputes to IMP_GP_TEMP and create up to 3 new impute variables using the following hierarchy:

i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:

1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_GRK_UNIT**= IMP_GP_TEMP

ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:

1. assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
2. set ALREADY_IMPUTED=1
3. Assign **MEDGP_GRK_ST**= IMP_GP_TEMP

iii. Otherwise:

1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_GRK**=IMP_GP_TEMP

h. Using **GQ_COUNTS_ROOMCAP_GREEK**, by UNITID, create unit-level sum variables (where a unit corresponds to a single UNITID, which corresponds to a single a university or college)

   i. Create unit-level sums (i.e., by UNITID) of GQCURRMAXPOP using only observations where flagD in (‘’,’R’). Note: these are the “good” values of GQCURRMAXPOP. Note that for this sum, we don’t care what the value of GP is, even it is a true 0. We are just trying to come up with a maximum number of people that these GQs *could* house, so that we can subtract the sum from the college-level IPEDS ROOMCAP variable. For reference later in the spec, call this sum **UNIT_MAXPOP_SUM**.

   ii. Using only the GQs with unres=0~~ and flagD **not** in (‘’,’R’)~~ and flagA not in ('I',’S’) and flagB not in ('I',’S’) and flagC not in ('I',’S’) and flagD = 'M' and GQCURRMAXPOP=., by UNITID, create unit-level sums of GP. Call this sum **UNIT_2020POP_SUM**.

   iii. Using only the GQs with~~ unres=1 and flagD **not** in (‘’,’R’)~~ (unres=1 or flagA = 'I' or flagB='I' or flagC='I' or flagD='I') and already_imputed=1 and GQCURRMAXPOP=., by UNITID, create unit-level sums of IMP_GP_TEMP. Call this **UNIT_POP_IMPUTED_SUM**.

   iv. Create **UNIT_CAP_SUM** = the unit-level sum of UNIT_MAXPOP_SUM, UNIT_2020POP_SUM, and UNIT_POP_IMPUTED_SUM

i. For each MAFID, calculate UNIT_RESIDUAL = ROOMCAP – UNIT_CAP_SUM (this will be the same value for MAFIDs with the same UNITID)

j. For each MAFID with UNIT_RESIDUAL<=0, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP, and create 3 new (non-Greek) median impute variables using the following hierarchy:

   i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:
      1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
      2. Set ALREADY_IMPUTED=1
      3. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP

   ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:
      1. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
      2. Set ALREADY_IMPUTED=1
      3. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP

   iii. Otherwise:
      1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
      2. Set ALREADY_IMPUTED=1
      3. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

k. For each (non-missing) UNITID with UNIT_RESIDUAL>0, count the MAFIDs associated with that UNITID that have UNRES=1 and ALREADY_IMPUTED=0. Call this count UNIT_RESID_GQ_COUNT.

l. For MAFIDs with UNIT_RESIDUAL>0, UNIT_RESID_GQ_COUNT=1, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy:

   i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK, then assign values to IMP_GP_TEMP using the following sub-hierarchy:

Commented [JEZ(F1]: Ask Kirk. This is like, if GQCURRMAXPOP is good, take that. Then if it’s flagged but the GQ is resolved, take GP (so this includes suppressed). Then if it’s unresolved, take imputed value. Some all of these to get a POP for the unit?

Commented [JEZ(F2R1]: This might change.

1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:
    - a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
    - b. Set ALREADY_IMPUTED=1
    - c. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP
2. Otherwise (i.e., if P50_GP_UNIT_BY_GRK<=0 or missing), if MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing, then:
    - a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
    - b. Set ALREADY_IMPUTED=1
    - c. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP
3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:
    - a. Assign IMP_GP_TEMP= P50_GP_BY_GRK
    - b. Set ALREADY_IMPUTED=1
    - c. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL < MAX_GP_UNIT_BY_GRK, then assign values as follows:

1. Assign IMP_GP_TEMP=UNIT_RESIDUAL
2. Set ALREADY_IMPUTED=1
3. Assign **IMP_RESID_1GQ**=IMP_GP_TEMP

m. For MAFIDs with UNIT_RESIDUAL>0, UNIT_RESID_GQ_COUNT>1, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy. (NOTE: steps i.1-i.3 are the same as steps i.1-i.3 in step l above):

i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK, then assign values to IMP_GP_TEMP using the following sub-hierarchy:

1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:
    - a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
    - b. Set ALREADY_IMPUTED=1
    - c. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP
2. Otherwise (i.e., if P50_GP_UNIT_BY_GRK<=0 or missing), if MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing, then:
    - a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
    - b. Set ALREADY_IMPUTED=1
    - c. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP
3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:
    - a. Assign IMP_GP_TEMP= P50_GP_BY_GRK
    - b. Set ALREADY_IMPUTED=1
    - c. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL < MAX_GP_UNIT_BY_GRK, then assign values as follows:

1. Assign IMP_GP_TEMP=UNIT_RESIDUAL/UNIT_RESID_GQ_COUNT
2. Set ALREADY_IMPUTED=1
3. Assign **IMP_RESID_NGQ**=IMP_GP_TEMP

n. Do a cross-tabulation of the variables UNRES and ALREADY_IMPUTED. If ALREADY_IMPUTED is always 1 when UNRES=1, then imputations have been calculated for all MAFIDS with GQCURTYP 501.

o. Keep the variables **MEDGP_GRK_UNIT, MEDGP_GRK_ST, MEDGP_GRK, MEDGP_nonGRK_UNIT, MEDGP_nonGRK_ST, MEDGP_nonGRK, IMP_RESID_1GQ**, and **IMP_RESID_NGQ.** Drop all other variables created in this section

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table hierarchically as follows, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. If IMP_RAT_EXP_GQ_ST is missing, if is not missing, assign IMP_GP = IMP_RAT_EXP_GQ and assign IMP_FLAG = 102. Continue on through the table until all MAFIDs with UNRES = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	IMP_FLAG
IMP_RAT_EXP_GQ_ST	101
IMP_RAT_EXP_GQ	102
IMP_RAT_EXP	103
IMP_RAT_MAX_GQ_ST	104
IMP_RAT_MAX_GQ	105
IMP_RAT_MAX	106
IMP_RAT_CURR_GQ_ST	107
IMP_RAT_CURR_GQ	108
IMP_RAT_CURR	109
IMP_RAT_CURRMAX_GQ_ST	110
IMP_RAT_CURRMAX_GQ	111
IMP_RAT_CURRMAX	112
MEDGP_GRK_UNIT	301
MEDGP_GRK_ST	302
MEDGP_GRK	303
MEDGP_nonGRK_UNIT	304
MEDGP_nonGRK_ST	305
MEDGP_nonGRK	306
IMP_RESID_1GQ	307
IMP_RESID_NGQ	308
IMP_MEDGP_GQ_ST	401
IMP_MEDGP_GQ	402
IMP_MEDGP	403

Section 6: Create Output Files

Output the following variables from GQMAFID:

Commented [JEZ(F3]: Ryan's recent files don't have geography on them…

MAFID	ACOCE	BCUCOUNTYFP
BCUSTATEFP	FACTLNAME	GQ_SIZE_EXP_PERS_CNT

GQ_SIZE_MAX_PERS_CNT	GQCONTACT	GQCURRMAXPOP
GQCURRSIZE	GQNAME	GQTYPCUR
GQ_INITIAL_STATUS	GQ_INITIAL_UNRES	GQ_INITIAL_POP
IMPUTE_NEEDED	FLAGA	FLAGB
FLAGC	FLAGD	GP
UNRES	IMP_GP	IMP_FLAG
EXPRATIO	EXPRATIO_GQ	EXPRATIO_GQ_ST
EXP_PERS_10	EXP_PERS_90	EXP_PERS_TRUNC
IMP_RAT_EXP	IMP_RAT_EXP_GQ	IMP_RAT_EXP_GQ_ST
MAXRATIO	MAXRATIO_GQ	MAXRATIO_GQ_ST
MAX_PERS_10	MAX_PERS_90	MAX_PERS_TRUNC
IMP_RAT_MAX	IMP_RAT_MAX_GQ	IMP_RAT_MAX_GQ_ST
CURRRATIO	CURRRATIO_GQ	CURRATIO_GQ_ST
CURRSIZE_10	CURRSIZE_90	CURRSIZE_TRUNC
IMP_RAT_CURR	IMP_RAT_CURR_GQ	IMP_RAT_CURR_GQ_ST
MAXCURRRATIO	MAXCURRRATIO_GQ	MAXCURRRATIO_GQ_ST
CURRMAX_10	CURRMAX_90	CURRMAX_TRUNC
IMP_RAT_CURRMAX	IMP_RAT_CURRMAX_GQ	IMP_RAT_CURRMAX_GQ_ST
MEDGP	MEDGP_GQ	MEDGP_GQ_ST
IMP_MEDGP	IMP_MEDGP_GQ	IMP_MEDGP_GQ_ST
MEDGP_GRK_UNIT	MEDGP_GRK_ST	MEDGP_GRK
MED_GP_nonGRK_UNIT	MEDGP_nonGRK_ST	MEDGP_nonGRK
IMP_RESID1GQ	IMP_RESID_NGQ	

Name this file gq_mafid_dssd_out_validation.sas7bdat

Output the following variables from GQMAFID:

MAFID	ACOCE	BCUCOUNTYFP
BCUSTATEFP	FACTLNAME	GQ_SIZE_EXP_PERS_CNT
GQ_SIZE_MAX_PERS_CNT	GQCONTACT	GQCURRMAXPOP
GQCURRSIZE	GQNAME	GQTYPCUR
GQ_INITIAL_STATUS	GQ_INITIAL_UNRES	GQ_INITIAL_POP
IMPUTE_NEEDED	FLAGA	FLAGB
FLAGC	FLAGD	GP
UNRES	IMP_GP	IMP_FLAG

Name this file gq_mafid_dssd_out_pop.sas7bdat. See POP data dictionary.

Andrew Keller, Julianne Zamora, Tim Kennel, Kirk White
December 26, 2020
2020 Census Specification For Group Quarters Imputation

Introduction

The goal of this specification is to impute occupied Group Quarters (GQ) as of April 1, 2020. This is necessary because some GQs were determined to be occupied on Census Day but a population count was unable to be obtained. The input file is total the GQ universe with population counts obtained via normal GQ operations (need to list) and a residual GQ call-in operation that occurred during December 2020. This input file has been created within DSSD. The output file is a list of MAFIDs that are to be imputed GQ counts. These counts will be assigned after the Decennial Response File 2 (DRF2) has been produced.

To summarize, the 2020 GQ Imputation specification is split into six sections:

1. Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation
2. Running HB Edits
3. Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation
4. Creating Imputed Values
5. Apply Ordering to Select Final Imputed Value
6. Create Output File

Input Files:

1. /sampling/eb/kelle321/gq_mafid_cnts_121920_geo_cdl2.sas7bdat (GQ_MAFID)
2. /sampling/share/hbparm.sas7bdat (HBPARM)
3. /sampling/share/gqmafid_undup_12220_more.sas7bdat (GQ_DUP_MAFID)
4. /sampling/share/mafid_frat_soro.csv (MAFID_FRAT_SORO)
5. /sampling/share/unitid_mafid_links.sas7bdat (UNITID_MAFID_LINKS)

Output Files: DSSD GQ Imputation Validation File (gq_mafid_dssd_out_validation.sas7bdat)
DSSD GQ Imputation Review File for POP (gq_mafid_dssd_out_pop.sas7bdat)

Section 1: Defining the Unresolved (Zero Pop) Cases Eligible for GQ Size Imputation

This section is divided into two steps. First, we must determine an initial pop count for resolved GQs those eligible for imputation because they are unresolved. Second, we use flags to determine outliers and put them into the imputation universe.

A. Ingest the input file (gq_mafid_cnts_121920_geo_cdl2.sas7bdat), referred to as ***GQ_MAFID***.
B. On this file, GQ_INITIAL_UNRES = 1 indicates an unresolved (zero pop) GQ
C. GQ_INITIAL_POP is the reported population before HB edits and imputation.
D. Rename GQ_INITIAL_POP to GQ_PRE_POP.

Section 1B: Reading in the Duplication Universe and Deducting Counts.

A. Ingest the input file (gqmafid_undup_12220_more.sas7bdat), referred to as ***GQ_DUP_MAFID***, keep only MAFID and SUM_GP_UNDUP.
B. Merge it to ***GQ_MAFID***, keeping all records in ***GQ_MAFID.***
C. Assign GQ_INITIAL_POP=GQ_PRE_POP.
D. If SUM_GP_UNDUP > 0 and SUM_GP_UNDUP < GQ_PRE_POP
    a. assign GQ_INITIAL_POP = SUM_GP_UNDUP.

<u>Section 2: HB Edits</u>

A. Calculate Ratios for editing.

  a. For each MAFID on ***GQ_MAFID***, if FOCS_ER_CB_CODE in (‘O’,’R’,‘ ‘) and GQ_INITIAL_POP > 0, then

    i. Assign **RATIOA** = GQ_INITIAL_POP/GQ_SIZE_EXP_PERS_CNT

    ii. Assign **RATIOB** = GQ_INITIAL_POP/GQ_SIZE_MAX_PERS_CNT

    iii. Assign **RATIOC** = GQ_INITIAL_POP/GQCURRSIZE

    iv. Assign **RATIOD** = GQ_INITIAL_POP/GQCURRMAXPOP

  b. Otherwise, RATIO[X] should be set to missing.

B. Create HB Parameters.

  a. For each MAFID on ***GQ_MAFID***, assign **GQTYPE** = first-digit of GQTYPCUR

  b. Read in parameters **C1**, **C2**, and **C3** for each RATIO[X] and GQTYPE on *HBPARM* (hbparm.sas7bdat) file.

GQTYPE	RATIO	C1	C2	C3
1	A	75	100	150
2	A	30	75	125
3	A	75	100	125
4	A	50	75	125
5	A	75	100	175
6	A	25	50	100
7	A	25	50	100
8	A	75	100	125
9	A	75	125	200
1	B	75	100	150
2	B	25	50	100
3	B	100	125	175
4	B	25	50	100
5	B	100	150	200
6	B	25	50	100
7	B	50	100	150
8	B	100	150	175
9	B	75	100	150
1	C	50	75	125
2	C	25	50	100
3	C	75	100	125
4	C	25	50	100
5	C	100	125	175
6	C	25	50	100
7	C	25	50	100
8	C	60	75	125
9	C	25	50	100
1	D	25	50	150
2	D	25	50	100
3	D	75	100	175

4	D	25	50	100
5	D	100	125	175
6	D	25	50	100
7	D	50	75	150
8	D	100	125	175
9	D	25	50	100

C. Run HB Edits for RATIOA, RATIOB, RATIOC, and RATIOD.

a. Apply steps b through h to each ratio separately. Calculate all quantiles and bounds by GQTYPE.

b. Merge the values of C1, C2, and C3 onto the ***GQ_MAFID*** file by merging HBPARM with ***GQ_MAFID*** file by GQTYPE for the given RATIO[X] X = A, B, C, or D.

c. For each GQTYPE, calculate the median value for RATIO[X] and assign this value as **MEDRATIO**.

d. For each MAFID, transform the ratio to create **SVALUE**.

i. If 0 < RATIO[X] < MEDRATIO then SVALUE = 1 – (MEDRATIO/RATIO[X])

ii. Else if RATIO[X] ≥ MEDRATIO then SVALUE = (RATIO[X]/MEDRATIO) - 1

e. For each MAFID, transform SVALUE to create **EVALUE**.

i. Calculate MAX_INTIAL_POP as max {GQ_INITIAL_POP, GQ_INITIAL_POP/RATIO[X]}

ii. Note, the second term in the brackets is the denominator of the RATIO[X] as GQ_INITIAL_POP is the numerator for all 4 ratios.

iii. EVALUE = SVALUE *(MAX_INITIAL_POP) $^{1/2}$

f. For each GQTYPE, calculate the first quartile, median, and third quartile of the EVALUE.

i. **E_Q1** = first quartile EVALUE

ii. **E_MED** = median EVALUE

iii. **E_Q3** = third quartile EVALUE

g. For each GQTYPE, define upper and lower bounds.

i. **D_Q1** = max {E_MED – E_Q1, abs (0.05*E_MED)}

ii. **D_Q3** = max {E_Q3 – E_MED, abs (0.05*E_MED)}

iii. **LOWER_C1** = E_MED – C1 * D_Q1

iv. **LOWER_C2** = E_MED – C2 * D_Q1

v. **LOWER_C3** = E_MED – C3 * D_Q1

vi. **UPPER_C1** = E_MED + C1 * D_Q3

vii. **UPPER_C2** = E_MED + C2 * D_Q3

viii. **UPPER_C3** = E_MED + C3 * D_Q3

h. For each MAFID, create **FLAG[X]**.

i. If EVALUE is missing, FLAG[X] = ‘M’

ii. Otherwise, apply the following conditions, without nesting (i.e. apply each ‘if’ statement separately).

1. If (EVALUE ≤ LOWER_C1 and LOWER_C1 is not missing) or (EVALUE ≥ UPPER_C1 AND UPPER_C1 not missing AND UPPER_C1 not equal to zero) then FLAG[X] = ‘R’

2. If (EVALUE ≤ LOWER_C2 and LOWER_C2 is not missing) or (EVALUE ≥ UPPER_C2 AND UPPER_C2 not missing AND UPPER_C2 not equal to zero) then FLAG[X] = 'S'
3. If (EVALUE ≤ LOWER_C3 and LOWER_C3 is not missing) or (EVALUE ≥ UPPER_C3 AND UPPER_C3 not missing AND UPPER_C3 not equal to zero) then FLAG[X] = 'I'

D. Update HB Flags for reasonable values of GQ_INITIAL_POP.

a. For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_INITIAL_POP for MAFIDs where FOCS_ER_CB_CODE in (' ','O','R') and GQ_INITIAL_UNRES = 0 and FLAGA not in ('S','I') and FLAGB not in ('S','I') and FLAGC not in ('S','I') and FLAGD not in ('S','I') and GQ_INITIAL_POP > 0. Assign these values as **GP_10** and **GP_90** respectively.

b. For each MAFID and FLAG[X] make the following update:

i. If FLAG[X] = 'I' and GQ_INITIAL_POP > GP_10 and GQ_INITIAL_POP < GP_90 then set FLAG[X] = 'S'.

c. For each MAFID, make the following updates:

i. If FLAGA = ' ' and FLAGB = 'I' then:

1. Set FLAGB = 'S'
2. If FLAGC = 'I' then set FLAGC = 'S'.
3. If FLAGD = 'I' then set FLAGD = 'S'.

ii. If FLAGA = ' ' and FLAGB = ' ' and FLAGC = 'I' then set FLAGC = 'S'.

~~i.~~iii. If FLAGA = ' ' and FLAGB = ' ' and FLAGC = ' ' and FLAGD = 'I' then set FLAGD = 'S'.

Commented [JEZ(F1]: Issue #2 with FLAGA and FLAGB = ' ' and FLAGC = 'I'.

E. Add flags FLAGA, FLAGB, FLAGC, and FLAGD onto ***GQ_MAFID***. All other variables created in this section should be dropped.

Section 3: Defining the Unresolved (Implausible Pop) Cases Eligible for GQ Size Imputation

A. After making initial determinations on what is eligible for imputation, we must remove outliers. These are initially resolved cases for which the result seems to be inconsistent with expectations. After this step, we will have our final universe for GQ imputation. The following variables will be assigned.

a. If (FLAGA = 'I' or FLAGB = 'I' or FLAGC = 'I' or FLAGD = 'I') and IMPUTE_NEEDED ne 'N' then

i. **GP = .**

ii. **UNRES** = 1

b. If MAFID = 'XXXXXXXXX' then set GP = . and UNRES = 1. Obtain MAFID from GQCI team.

Commented [JEZ(F2]: Issue #1 with all reported pop in one dorm.

c. Otherwise,

i. **GP** = GQ_INITIAL_POP

ii. **UNRES** = GQ_INITIAL_UNRES

Section 4: Create Imputed Values

This section develops the Imputation Models Estimation Methodology for multiple imputation methods to choose between depending on the GQ type. There are four imputation subsections that document the various imputation methods.

A. Assign Ratio-Adjustment Values

a. Calculate GP/GQ_SIZE_EXP_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.
We will create 3 ratios comparing GP to GQ_SIZE_EXP_PERS_CNT, one for the national value (**EXPRATIO**), one for the GQTYPCUR combination (**EXPRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**EXPRATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’):

1. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for the nation.**
2. Assign **EXPRATIO** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each GQTYPCUR value.**
4. Assign **EXPRATIO_GQ** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID
5. Sum the GP and GQ_SIZE_EXP_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **EXPRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_EXP_PERS_CNT) for each MAFID.

ii. Calculate Bounds.
For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_EXP_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGA in (‘ ‘,’R’). Assign these values as **EXP_PERS_10** and **EXP_PERS_90** respectively.
For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_EXP_PERS_CNT.

1. Assign **EXP_PERS_TRUNC** = GQ_SIZE_EXP_PERS_CNT
2. If GQ_SIZE_EXP_PERS_CNT > EXP_PERS_90 then set **EXP_PERS_TRUNC** = EXP_PERS_90
3. If GQ_SIZE_EXP_PERS_CNT > 0 and GQ_SIZE_EXP_PERS_CNT < EXP_PERS_10 then set **EXP_PERS_TRUNC** = EXP_PERS_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_EXP** = CEIL (EXP_PERS_TRUNC*EXPRATIO)
2. **IMP_RAT_EXP_GQ** = CEIL (EXP_PERS_TRUNC*EXPRATIO_GQ)
3. **IMP_RAT_EXP_GQ_ST** = CEIL (EXP_PERS_TRUNC*EXPRATIO_GQ_ST)

b. Calculate GP/GQ_SIZE_MAX_PERS_CNT Ratio-Adjusted Imputed Values

i. Calculate Ratios.
We will create 3 ratios comparing GP to GQ_SIZE_MAX_PERS_CNT, one for the national value (**MAXRATIO**), one for the GQTYPCUR combination (**MAXRATIO_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagB in (‘’,’R’):

1. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for the nation.**
2. Assign **MAXRATIO** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.
3. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each GQTYPCUR value.**
4. Assign **MAXRATIO_GQ** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID

5. Sum the GP and GQ_SIZE_MAX_PERS_CNT value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **MAXRATIO_GQ_ST** = sum(GP)/sum(GQ_SIZE_MAX_PERS_CNT) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQ_SIZE_MAX_PERS_CNT for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGB in (‘ ‘,’R’). Assign these values as **MAX_PERS_10** and **MAX_PERS_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQ_SIZE_MAX_PERS_CNT.

1. Assign **MAX_PERS_TRUNC** = GQ_SIZE_MAX_PERS_CNT
2. If GQ_SIZE_MAX_PERS_CNT > MAX_PERS_90 then set **MAX_PERS_TRUNC** = MAX_PERS_90
3. If GQ_SIZE_MAX_PERS_CNT > 0 and GQ_SIZE_MAX_PERS_CNT < MAX_PERS_10 then set **MAX_PERS_TRUNC** = MAX_PERS_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_MAX** = CEIL (MAX_PERS_TRUNC*MAXRATIO)
2. **IMP_RAT_MAX_GQ** = CEIL (MAX_PERS_TRUNC*MAXRATIO_GQ)
3. **IMPRAT_MAX_GQ_ST** = CEIL (MAX_PERS_TRUNC*MAXRATIO_GQ_ST)

c. Calculate GP/GQCURRSIZE Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRSIZE, one for the national value (**CURRSIZERATIO)**, one for the GQTYPCUR combination (**CURRSIZERATIO_GQ)**, and one for the GQTYPCUR and BCUSTATEFP combination (**CURRSIZERATIO_GQ_ST)**. If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagC in (‘’,’R’):

1. Sum the GP and GQCURRSIZE value **for the nation.**
2. Assign **CURRSIZERATIO** = sum(GP)/sum(GQCURRSIZE) for each MAFID.
3. Sum the GP and GQCURRSIZE value **for each GQTYPCUR value.**
4. Assign **CURRSIZERATIO_GQ** = sum(GP)/sum(GQCURRSIZE) for each MAFID
5. Sum the GP and GQCURRSIZE value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRSIZERATIO_GQ_ST** = sum(GP)/sum(GQCURRSIZE) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRSIZE for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGC in (‘ ‘,’R’). Assign these values as **CURRSIZE_10** and **CURRSIZE_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRSIZE.

1. Assign **CURRSIZE_TRUNC** = GQCURRSIZE
2. If GQCURRSIZE > CURRSIZE_90 then set **CURRSIZE_TRUNC** = CURRSIZE_90

3. If GQCURRSIZE > 0 and GQCURRSIZE < CURRSIZE_10 then set **CURRSIZE_TRUNC** = CURRSIZE_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURR** = CEIL (CURRSIZE_TRUNC*CURRSIZERATIO)
2. **IMP_RAT_CURR_GQ** = CEIL (CURRSIZE_TRUNC*CURRSIZERATIO_GQ)
3. **IMP_RAT_CURR_GQ_ST** = CEIL (CURRSIZE_TRUNC*CURRSIZERATIO_GQ_ST)

d. Calculate GP/GQCURRMAXPOP Ratio-Adjusted Imputed Values

i. Calculate Ratios.

We will create 3 ratios comparing GP to GQCURRMAXPOP, one for the national value (**CURRMAXRATIO**), one for the GQTYPCUR combination (**CURRMAXRATIO_GQ**), and one for the GQTYPCUR nd BCUSTATEFP combination (**CURRMAXRATIO_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagD in (‘’,’R’):

1. Sum the GP and GQCURRMAXPOP value **for the nation.**
2. Assign **CURRMAXRATIO** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.
3. Sum the GP and GQCURRMAXPOP value **for each GQTYPCUR value.**
4. Assign **CURRMAXRATIO_GQ** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID
5. Sum the GP and GQCURRMAXPOP value **for each combination of GQTYPCUR and BCUSTATEFP value.**
6. Assign **CURRMAXRATIO_GQ_ST** = sum(GP)/sum(GQCURRMAXPOP) for each MAFID.

ii. Calculate Bounds.

For each GQTYPCUR, calculate the 10th and 90th percentiles of GQCURRMAXPOP for MAFIDs where UNRES = 0 and FOCS_ER_CB_CODE = ‘ ‘ and FLAGD in (‘ ‘,’R’). Assign these values as **CURRMAX_10** and **CURRMAX_90** respectively.

For each MAFID where UNRES = 1 , assign truncated values of GQCURRMAXPOP.

1. Assign **CURRMAX_TRUNC** = GQCURRMAXPOP
2. If GQCURRMAXPOP > CURRMAX_90 then set **CURRMAX_TRUNC** = CURRMAX_90
3. If GQCURRMAXPOP > 0 and GQCURRMAXPOP < CURRMAX_10 then set **CURRMAX_TRUNC** = CURRMAX_10.

iii. Assign values. For each MAFID, calculate the following values:

1. **IMP_RAT_CURRMAX** = CEIL (CURRMAX_TRUNC*CURRMAXRATIO)
2. **IMP_RAT_CURRMAX_GQ** = CEIL (CURRMAX_TRUNC*CURRMAXRATIO_GQ)
3. **IMP_RAT_CURRMAX_GQ_ST** = CEIL (CURRMAX_TRUNC*CURRMAXRATIO_GQ_ST)

B. Assign Good Person Percentile counts.

a. We will create 3 Good Person Percentile counts, one for the national value (**MEDGP**), one for the GQTYPCUR combination (**MEDGP_GQ**), and one for the GQTYPCUR and BCUSTATEFP combination (**MEDGP_GQ_ST**). If unres = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’,’R’):

i. Find the 65th percentile on GP **for the nation.** Assign it as **MEDGP.**

ii. Find the 65th percentile on GP **for each GQTYPCUR value.** Assign them as **MEDGP_GQ.**

iii. Find the 65th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

1. For GQTYPCUR=104, 801, 802, 901 find the 70th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
2. For GQTYPCUR=501 find the 68th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Overwrite them as **MEDGP_GQ_ST.**
3. For GQTYPCUR=301, find the 55th percentile on GP **for each combination of GQTYPCUR and BCUSTATEFP value.** Assign them as **MEDGP_GQ_ST.**

iv. Assign values. For each MAFID, calculate the following values:

1. **IMP_MEDGP_GQ_ST** = CEIL(MEDGP_GQ_ST)
2. **IMP_MEDGP_GQ** = CEIL(MEDGP_GQ)
3. **IMP_MEDGP** = CEIL(MEDGP)

C. CES method: impute using a hybrid of the ratio imputes created in the previous step, a percentile method based on Greek/non-Greek status, and a facility-level residual allocation method.

a. Ingest the file referred to as **MAFID_FRAT_SORO**

i. On this file **FLAG_GREEK_LETTER**=1 indicates that GQ has been identified as a fraternity or sorority house. Otherwise **FLAG_GREEK_LETTER**=0.

b. Ingest the file referred to as **UNITID_MAFID_LINKS**.

i. When reading in **UNITID_MAFID_LINKS,** keep only the variables **MAFID, UNITID, MATCH_STEP_NUM**, and **ROOMCAP.**

ii. Note: for records with **MATCH_STEP_NUM**=-1, **UNITID** will be missing.

iii. Note: for records with the same value of UNITID, ROOMCAP will be the same.

c. Merge **MAFID_FRAT_SORO** and **UNITID_MAFID_LINKS** to ***GQ_MAFID***, merging on MAFID, and keeping only records that are in ***GQ_MAFID.***

i. Note: For records that match, this should be a 1-to-1 match (MAFID should be unique in each of the 3 datasets).

ii. Note: only records with GQCURTYP=501 in ***GQ_MAFID*** should match to either of the other 2 datasets.

d. Select the subset of the merged dataset from the previous step with GQCURTYP=501.

i. NOTE: In this spec we will refer to this subset of the data as **GQ_COUNTS_ROOMCAP_GREEK**. This is only an intermediate dataset, which will be merged back to the **GQ_MAFID** dataset at the end of this section of the spec (section 5.D).

e. Using GQ_COUNTS_ROOM_CAP_GREEK and the ratio impute variables created in section 4.A, create a temporary impute variable IMP_GP_TEMP using the hierarchy shown in the following table. If IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP_TEMP= IMP_RAT_EXP_GQ_ST and set ALREADY_IMPUTED=1. If IMP_RAT_EXP_GQ_ST is missing and IMP_RAT_EXP_GQ is not missing, assign IMP_GP_TEMP= IMP_RAT_EXP_GQ and set ALREADY_IMPUTED=1. Continue through the table until all the variables in the table have been exhausted. For any remaining

MAFIDs for which a value has not been assigned to IMP_GP_TEMP, set ALREADY_IMPUTED=0;

IMP_GP_TEMP assignment hierarchy
IMP_RAT_EXP_GQ_ST
IMP_RAT_EXP_GQ
IMP_RAT_MAX_GQ_ST
IMP_RAT_MAX_GQ
IMP_RAT_CURR_GQ_ST
IMP_RAT_CURR_GQ
IMP_RAT_CURRMAX_GQ_ST
IMP_RAT_CURRMAX_GQ

f. Using only MAFIDs in **GQ_COUNTS_ROOMCAP_GREEK** with UNRES = 0 and FOCS_ER_CB_CODE = ‘’ and flagA in (‘’,’R’) and flagB in (‘’,’R’) and flagC in (‘’,’R’) and flagD in (‘’,’R’), create 3 GP median variables and 3 GP maximum variables:
   i. For each UNITID-FLAG_GREEK_LETTER combination:
      1. Calculate the median value of GP. Call this **P50_GP_UNIT_BY_GRK**
      2. Calculate the maximum value of GP. Call this **MAX_GP_UNIT_BY_GRK.**
      3. Merge the P50_GP_UNIT_BY_GRK and MAX_GP_UNIT_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on UNITID and FLAG_GREEK_LETTER.
      4. Note, these values will be missing if there are not enough observations for the UNITID-FLAG_GREEK_LETTER combination.
   ii. For each BCUSTATEFP-FLAG_GREEK_LETTER combination:
      1. Calculate the median value of GP. Call this **P50_GP_ST_BY_GRK**.
      2. Calculate the maximum value of GP. Call this **MAX_GP_ST_BY_GRK**.
      3. Merge P50_GP_ST_BY_GRK and MAX_GP_ST_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on BCUSTATEFP-FLAG_GREEK_LETTER combinations.
      4. Note, these values will be missing if there are not enough observations for the BCUSTATEFP-FLAG_GREEK_LETTER combination.
   iii. For each value of FLAG_GREEK_LETTER:
      1. Calculate the median value of GP. Call this **P50_GP_BY_GRK.**
      2. Calculate the maximum value of GP. Call this **MAX_GP_BY_GRK**.
      3. Merge P50_GP_BY_GRK and MAX_BP_BY_GRK back onto **GQ_COUNTS_ROOMCAP_GREEK**, merging on FLAG_GREEK_LETTER.

g. For MAFIDs for which UNRES=1, FLAG_GREEK_LETTER=1, and ALREADY_IMPUTED=0, assign median Greek imputes to IMP_GP_TEMP and create up to 3 new impute variables using the following hierarchy:
   i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:
      1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
      2. Set ALREADY_IMPUTED=1
      3. Assign **MEDGP_GRK_UNIT**= IMP_GP_TEMP
   ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:
      1. assign IMP_GP_TEMP= P50_GP_ST_BY_GRK

2. set ALREADY_IMPUTED=1
3. Assign **MEDGP_GRK_ST**= IMP_GP_TEMP

iii. Otherwise:

1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_GRK**=IMP_GP_TEMP

h. Using **GQ_COUNTS_ROOMCAP_GREEK**, by UNITID, create unit-level sum variables (where a unit corresponds to a single UNITID, which corresponds to a single a university or college)

i. Create unit-level sums (i.e., by UNITID) of GQCURRMAXPOP using only observations where flagD in (‘’,’R’). Note: these are the “good” values of GQCURRMAXPOP. Note that for this sum, we don’t care what the value of GP is, even it is a true 0. We are just trying to come up with a maximum number of people that these GQs *could* house, so that we can subtract the sum from the college-level IPEDS ROOMCAP variable. For reference later in the spec, call this sum **UNIT_MAXPOP_SUM**.

ii. Using only the GQs with unres=0 and flagA not in ('I',’S’) and flagB not in ('I',’S’) and flagC not in ('I',’S’) and flagD = 'M' and GQCURRMAXPOP=., by UNITID, create unit-level sums of GP. Call this sum **UNIT_2020POP_SUM**.

iii. Using only the GQs with (unres=1 or flagA = 'I' or flagB='I' or flagC='I' or flagD='I') and already_imputed=1 and GQCURRMAXPOP=., by UNITID, create unit-level sums of IMP_GP_TEMP. Call this **UNIT_POP_IMPUTED_SUM**.

iv. Create **UNIT_CAP_SUM** = the unit-level sum of UNIT_MAXPOP_SUM, UNIT_2020POP_SUM, and UNIT_POP_IMPUTED_SUM

i. For each MAFID, calculate UNIT_RESIDUAL = ROOMCAP – UNIT_CAP_SUM (this will be the same value for MAFIDs with the same UNITID)

j. For each MAFID with UNIT_RESIDUAL<=0, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP, and create 3 new (non-Greek) median impute variables using the following hierarchy:

i. If **P50_GP_UNIT_BY_GRK** >0 and not missing:

1. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP

ii. If **P50_GP_UNIT_BY_GRK** <=0 or missing and **P50_GP_ST_BY_GRK**>0 and not missing, then:

1. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP

iii. Otherwise:

1. Assign IMP_GP_TEMP= P50_GP_BY_GRK
2. Set ALREADY_IMPUTED=1
3. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

k. For each (non-missing) UNITID with UNIT_RESIDUAL>0, count the MAFIDs associated with that UNITID that have UNRES=1 and ALREADY_IMPUTED=0. Call this count UNIT_RESID_GQ_COUNT.

l. For MAFIDs with UNIT_RESIDUAL>0, UNIT_RESID_GQ_COUNT=1, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy:
  - i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK, then assign values to IMP_GP_TEMP using the following sub-hierarchy:
    - 1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:
      - a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
      - b. Set ALREADY_IMPUTED=1
      - c. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP
    - 2. Otherwise (i.e., if P50_GP_UNIT_BY_GRK<=0 or missing), if MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing, then:
      - a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
      - b. Set ALREADY_IMPUTED=1
      - c. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP
    - 3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:
      - a. Assign IMP_GP_TEMP= P50_GP_BY_GRK
      - b. Set ALREADY_IMPUTED=1
      - c. Assign **MEDGP_nonGRK**=IMP_GP_TEMP
  - ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL < MAX_GP_UNIT_BY_GRK, then assign values as follows:
    - 1. Assign IMP_GP_TEMP=UNIT_RESIDUAL
    - 2. Set ALREADY_IMPUTED=1
    - 3. Assign **IMP_RESID_1GQ**=IMP_GP_TEMP

m. For MAFIDs with UNIT_RESIDUAL>0, UNIT_RESID_GQ_COUNT>1, UNRES=1, and ALREADY_IMPUTED=0, assign values to IMP_GP_TEMP and ALREADY_IMPUTED and create (up to) 1 new impute variables using the following hierarchy. (NOTE: steps i.1-i.3 are the same as steps i.1-i.3 in step l above):
  - i. If MAX_GP_UNIT_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_UNIT_BY_GRK, then assign values to IMP_GP_TEMP using the following sub-hierarchy:
    - 1. If P50_GP_UNIT_BY_GRK>0 and non-missing, then:
      - a. Assign IMP_GP_TEMP= P50_GP_UNIT_BY_GRK
      - b. Set ALREADY_IMPUTED=1
      - c. Assign **MEDGP_nonGRK_UNIT**= IMP_GP_TEMP
    - 2. Otherwise (i.e., if P50_GP_UNIT_BY_GRK<=0 or missing), if MAX_GP_ST_BY_GRK>0 and non-missing and UNIT_RESIDUAL > MAX_GP_ST_BY_GRK and P50_GP_ST_BY_GRK>0 and non-missing, then:
      - a. Assign IMP_GP_TEMP= P50_GP_ST_BY_GRK
      - b. Set ALREADY_IMPUTED=1
      - c. Assign **MEDGP_nonGRK_ST**= IMP_GP_TEMP
    - 3. Otherwise (i.e., if the conditions in steps i. and ii. are not met), then:
      - a. Assign IMP_GP_TEMP= P50_GP_BY_GRK
      - b. Set ALREADY_IMPUTED=1
      - c. Assign **MEDGP_nonGRK**=IMP_GP_TEMP

   - ii. If MAX_GP_UNIT_BY_GRK=0 or missing or UNIT_RESIDUAL < MAX_GP_UNIT_BY_GRK, then assign values as follows:
     1. Assign IMP_GP_TEMP=UNIT_RESIDUAL/UNIT_RESID_GQ_COUNT
     2. Set ALREADY_IMPUTED=1
     3. Assign **IMP_RESID_NGQ**=IMP_GP_TEMP

n. Do a cross-tabulation of the variables UNRES and ALREADY_IMPUTED. If ALREADY_IMPUTED is always 1 when UNRES=1, then imputations have been calculated for all MAFIDS with GQCURTYP 501.

o. Keep the variables **MEDGP_GRK_UNIT, MEDGP_GRK_ST, MEDGP_GRK, MEDGP_nonGRK_UNIT, MEDGP_nonGRK_ST, MEDGP_nonGRK, IMP_RESID_1GQ**, and **IMP_RESID_NGQ.** Drop all other variables created in this section

Section 5: Apply Ordering to Select Final Imputed Value

For each MAFID where unres = 1, use the following table to assign the imputed value IMP_GP and IMP_FLAG. Read the table hierarchically as follows, if IMP_RAT_EXP_GQ_ST is not missing, assign IMP_GP = IMP_RAT_EXP_GQ_ST and assign IMP_FLAG = 101. If IMP_RAT_EXP_GQ_ST is missing, if IMP_RAT_EXP_GQ is not missing, assign IMP_GP = IMP_RAT_EXP_GQ and assign IMP_FLAG = 102. Continue on through the table until all MAFIDs with UNRES = 1 have a value for IMP_GP and IMP_FLAG.

IMP_GP	IMP_FLAG
IMP_RAT_EXP_GQ_ST	101
IMP_RAT_EXP_GQ	102
IMP_RAT_EXP	103
IMP_RAT_MAX_GQ_ST	104
IMP_RAT_MAX_GQ	105
IMP_RAT_MAX	106
IMP_RAT_CURR_GQ_ST	107
IMP_RAT_CURR_GQ	108
IMP_RAT_CURR	109
IMP_RAT_CURRMAX_GQ_ST	110
IMP_RAT_CURRMAX_GQ	111
IMP_RAT_CURRMAX	112
MEDGP_GRK_UNIT	301
MEDGP_GRK_ST	302
MEDGP_GRK	303
MEDGP_nonGRK_UNIT	304
MEDGP_nonGRK_ST	305
MEDGP_nonGRK	306
IMP_RESID_1GQ	307
IMP_RESID_NGQ	308
IMP_MEDGP_GQ_ST	401
IMP_MEDGP_GQ	402
IMP_MEDGP	403

Section 6: Create Output Files

Output the following variables from GQMAFID:

Commented [JEZ(F3]: Ryan's recent files don't have geography on them…

MAFID	ACOCE	BCUCOUNTYFP
BCUSTATEFP	FACTLNAME	GQ_SIZE_EXP_PERS_CNT
GQ_SIZE_MAX_PERS_CNT	GQCONTACT	GQCURRMAXPOP
GQCURRSIZE	GQNAME	GQTYPCUR
GQ_INITIAL_STATUS	GQ_INITIAL_UNRES	GQ_INITIAL_POP
IMPUTE_NEEDED	FLAGA	FLAGB
FLAGC	FLAGD	GP
UNRES	IMP_GP	IMP_FLAG
EXPRATIO	EXPRATIO_GQ	EXPRATIO_GQ_ST
EXP_PERS_10	EXP_PERS_90	EXP_PERS_TRUNC
IMP_RAT_EXP	IMP_RAT_EXP_GQ	IMP_RAT_EXP_GQ_ST
MAXRATIO	MAXRATIO_GQ	MAXRATIO_GQ_ST
MAX_PERS_10	MAX_PERS_90	MAX_PERS_TRUNC
IMP_RAT_MAX	IMP_RAT_MAX_GQ	IMP_RAT_MAX_GQ_ST
CURRRATIO	CURRRATIO_GQ	CURRATIO_GQ_ST
CURRSIZE_10	CURRSIZE_90	CURRSIZE_TRUNC
IMP_RAT_CURR	IMP_RAT_CURR_GQ	IMP_RAT_CURR_GQ_ST
CURRMAXRATIO	CURRMAXRATIO_GQ	CURRMAXRATIO_GQ_ST
CURRMAX_10	CURRMAX_90	CURRMAX_TRUNC
IMP_RAT_CURRMAX	IMP_RAT_CURRMAX_GQ	IMP_RAT_CURRMAX_GQ_ST
MEDGP	MEDGP_GQ	MEDGP_GQ_ST
IMP_MEDGP	IMP_MEDGP_GQ	IMP_MEDGP_GQ_ST
MEDGP_GRK_UNIT	MEDGP_GRK_ST	MEDGP_GRK
MEDGP_nonGRK_UNIT	MEDGP_nonGRK_ST	MEDGP_nonGRK
IMP_RESID_1GQ	IMP_RESID_NGQ	

Name this file gq_mafid_dssd_out_validation.sas7bdat

Output the following variables from GQMAFID:

MAFID	ACOCE	BCUCOUNTYFP
BCUSTATEFP	FACTLNAME	GQ_SIZE_EXP_PERS_CNT
GQ_SIZE_MAX_PERS_CNT	GQCONTACT	GQCURRMAXPOP
GQCURRSIZE	GQNAME	GQTYPCUR
GQ_INITIAL_STATUS	GQ_INITIAL_UNRES	GQ_INITIAL_POP
IMPUTE_NEEDED	FLAGA	FLAGB
FLAGC	FLAGD	GP
UNRES	IMP_GP	IMP_FLAG
CALL_STATUS	GEO_POP_COUNT	

Name this file gq_mafid_dssd_out_pop.sas7bdat. See POP data dictionary.

Update on Off Campus Housing Unit Records Collection

Thomas Mule

July 9, 2020

Pre-Decisional: Internal Use Only

DRB Approval Number: CBDRB-FY21-DSEP-002

Off Campus Housing Unit Records Collection

Census Bureau is contacting universities to see if they can provide information about their students who live in off-campus housing

- Non-Group Quarters population

Requesting universities provide:

- Student first name, middle name, last name, month of birth, day of birth, year of birth, age
  - Had been requesting sex, race and Hispanic origin (Y or N) but discontinued
- Local off-campus address
- Alternative address

DRB Approval Number: CBDRB-FY21-DSEP-002

Researching Three Usages

1. Can we use this information combined with other administrative records to enumerate the off-campus household?
   - Use during Closeout phase of NRFU
   - Use during Post processing
     - Off-campus roster is used instead of vacant NRFU interview
     - If roster is not complete enough, can we determine housing unit is occupied with unknown population count
2. Can the alternative address help identify duplication between the off campus enumeration and the alternative address enumeration?
3. Can the off-campus information be used to add the student to an off-campus enumeration
   - Incomplete self-response or NRFU enumeration did not include the student

Any production implementation would require system development and testing

Status of Data Collection

Universities or Colleges	Number
Universe	1365
Contacted	782
Participating	383
Not Participating	250
Files Received	112
Files Formatted	5
Files Sent from GEO to 2020 Usage Group	5

While Geography and DSSD were working out the automation to do the final transfers,

Geography Division has started to format, assign MAFIDs and geocode the addresses

DSSD has been looking at the raw received files and initial geography results

DRB Approval Number: CBDRB-FY21-DSEP-002

Initial Analysis of Received Files

- While we provided them a template, schools are able to submit their information in any form
  - Month, day and year of birth in one variable
  - Addresses information (house number, street, within structure, city, state and zip code) all in one field
  - Geography division is doing followup calls when needed

2020 Usage Group has started looking at 39 school files

- Next slides is some initial results that will change as the processing is continued to be refined.

Person Information

Analysis of 39 Schools

- 139,857 person records
  - [redacted]
- Seeing fairly complete reporting of first and last name
- 7 schools did not provide date of birth
- 7 schools provided sex
- 9 schools provided information in race field
- 7 schools provided information in Hispanic field

Local and Alternative Address Information

- 12 of the 39 schools only provided one address
  - Geography division is doing followup calls to confirm
- Some schools are not providing zip codes for all of the addresses
  - Can impact the assigning of MAFIDs to addresses

Research Steps

Geography Division and 2020 Usage Group

- Standardizing the person and address characteristics
- Assigning MAFIDs to the addresses
- Assigning geocodes (state, county, tract, block) if MAFID can not be assigned

2020 Usage Group

- Matching the off-campus responses against the Self-Response Quality Assurance (SRQA) composite
- Assigning SRQA administrative record person ID allows us to link to NRFU AR Modeling Input files and AR Person Lookup Characteristics
- Using this information to see how this data can be used for three research question to address coverage of this population

2020 Data Quality Executive Governance Group Meeting
December 3, 2020

EGG Members:

X	John Abowd
X	Pat Cantwell
X	Jamey Christy
X	Al Fontenot
X	Ron Jarmin
	Christa Jones
X	Enrique Lamas
X	Ben Page
X	Deb Stempowski
X	Tori Velkoff

X = Present

Others:

X	Jennifer Ortman
	Burton Reist
	Deirdre Bishop
	Mike Ratcliffe
X	Michael Thieme
X	Maryann Chapin
X	Jennifer Reichert
X	Karen Battle
X	Jon Spader
X	Christine Borman
X	Roberto Ramirez
X	Jason Devine
X	Marc Perry
	Stephanie Galvin

Agenda

1. Timing for Release of Table Set 3 (Persons Enumerated in Group Quarters)
2. Data Review Update (DEMO Staff)
3. External Engagements (JASON, ASA, CNSTAT, OIG, and GAO)

Meeting Minutes

Timing for Release of Table Set 3

This table would need a privacy loss budget if released in advance of the redistricting data. The EGG supported moving this table set to release alongside the redistricting data.

John and Pat will discuss offline resolving the DA issues in Table 2 with regard to the GQ population.

Supporting documentation for this item is saved at:

[redacted]

Data Review Update

DEMO staff provided an update on the data review that is in progress, highlighting some initial review findings.

The EGG discussion centered on initial findings from review of group quarters data in the DRF1 and next steps to resolve issues. The group will reconvene on 12/4 to discuss this matter further.

Supporting documentation for this item is saved at:

[redacted]

External Engagement

Updates provided via meeting minutes as there was not sufficient time to discuss at this week's EGG meeting.

JASON: work continues to get an agreement in place, work is expected to begin very soon. There is a meeting with the JASONs on Friday to get information about the type of briefings and information they would like to receive.

ASA: conversations continue to scope out and plan this collaboration, another meeting (Census + ASA) is scheduled for 12/4.

CNSTAT: No updates on a Census engagement with CNSTAT. CNSTAT is sponsoring a seminar focused around the ASA Task Force Document on 2020 Census Quality Indicators. The seminar may be occurring on December 16.

OIG: An initial discussion for the upcoming engagement on quality of the 2020 Census data took place on 12/1. OIG has requested:

1. *Near Term Deliverables*
   a. *Inventory of planned approaches and metrics to inform understanding of data quality*
   b. *Inventory or operation impacts that impact data quality*
2. *Long Term Deliverables-Approach to monitoring and assessing quality and communicating information to the Operational Update Team*
3. *Metrics that will be issued with the Apportionment and Redistricting*
4. *2010 "Data Quality Document"*

GAO: GAO is participating in a House Oversight hearing this Thursday (12/3).

2020 Data Quality Executive Governance Group Meeting
December 4, 2020

EGG Members:

X	John Abowd
X	Pat Cantwell
X	Jamey Christy
X	Al Fontenot
X	Ron Jarmin
X	Christa Jones
X	Enrique Lamas
X	Ben Page
X	Deb Stempowski
X	Tori Velkoff

X = Present

Others:

X	Jennifer Ortman
X	Burton Reist
X	Deirdre Bishop
X	Mike Ratcliffe
X	Michael Thieme
X	Maryann Chapin
X	Jennifer Reichert
X	Karen Battle
X	Jon Spader
X	Christine Borman
X	Roberto Ramirez
X	Jason Devine
X	Marc Perry
X	Louis Avenilla
X	Barbara LoPresti
X	Andrea Johnson
X	Judy Belton
X	Karen Field
X	Stuart Irby
X	Debbie Fenstermaker
X	Tom Mule
X	Andy Keller
X	Steve Wilson
X	Kin Koerber

Agenda

1. Data Review Update – continued discussion of GQ from 12/3 meeting

Meeting Minutes

The EGG continued the discussion that began on 12/3/20 about the initial data review findings for group quarters data.

Analysis continues to determine the magnitude of the issue. Teams from GEO, POP, and DSSD are working on this.

- GEO is looking at tract-level data to see where 2020 GQ populations do not align with the benchmark data. Their initial focus is to identify situations where the 2020 GQ population in DRF1 is *lower* than expected.
- DSSD will begin a record linkage exercise for the GQ universe to supplement the existing unduplication efforts to assess if additional measures for the unduplication within a facility are

needed. This will help to identify GQs where the DRF1 count is higher than the benchmark where duplication is suspected.

- POP will continue a first pass at tract-level maps of GQ population data compared with benchmarks, tabulating a list of GQ facilities by state with populations that differ substantially.

As soon as possible, but no later than Sunday night, the EGG would like to receive information indicating whether there is a potential impact to apportionment. This group will reconvene Sunday night at 7pm.

The EGG would like to see a list of GQs by type with zero population by state. Once that information is produced, additional thresholds should be established to show the range of GQs in different scenarios.

POCs:

- POP and SEHSD will identify a staff member to serve as coordinator for work being conducted across divisions.
- DSSD will also provide a POC for the coordination of this work.
- Shawn Klimek will serve as the POC within R&M.

2020 Data Quality Executive Governance Group Meeting
December 6, 2020

EGG Members:

X	John Abowd
X	Pat Cantwell
X	Jamey Christy
X	Al Fontenot
X	Ron Jarmin
X	Christa Jones
X	Enrique Lamas
X	Ben Page
X	Deb Stempowski
X	Tori Velkoff

X = Present

Others:

X	Jennifer Ortman
	Burton Reist
X	Deirdre Bishop
X	Mike Ratcliffe
	Michael Thieme
X	Maryann Chapin
X	Jennifer Reichert
X	Karen Battle

X	Andrea Johnson
X	Andy Keller
X	Barbara LoPresti
X	Christine Borman
X	Colleen Keating
X	Debbie Fenstermaker
X	Derek Breese
X	Jason Devine
X	Jon Spader
X	Judy Belton
X	Karen Field
X	Laura Wagoneer
X	Lauren Medina
X	Lindsay Spell
X	Louis Avenilla
X	Marc Perry
X	Matt Spence
X	Roberto Ramirez
X	Steve Wilson
X	Stuart Irby
X	Tom Mule
X	Kin Koerber

Agenda

1. Diagnosing the Problem - POP then GEO presents their findings and (preliminary) conclusions
2. Potential Actions - DSSD presents their findings
3. EGG Discussion

Meeting Materials

Meeting materials are saved at:

[redacted]

Research Findings:

POP and GEO presented the findings of their research. The work carried out this weekend indicates there are issues to address in the GQ data. More investigation is needed to fully understand the issue.

DSSD has been exploring options for count imputation of GQ population as well as the potential use of record linkage to identify duplicates.

The EGG asked for:

- A list of refusals.

- A list of 501s that did not respond at all.
- Information about the data used by GEO to calculate the summary measures by GQ type (i.e., what the numerators and denominators are).
- The number of GQs that do not have a value from the advance contact.
- The detail information behind the list of schools from the POP slide deck (send to Judy, Jennifer R., and Deb).

Jennifer O. will schedule a meeting for the teams to reconvene in the morning to discuss next steps.

2020 Data Quality Executive Governance Group Meeting

December 10, 2020

Participants

Al Fontenot
Andrea Johnson
Andy Keller
Barbara LoPresti
Ben Page
Burton Reist
Christa Jones
Deb Stempowski
Debbie Fenstermaker
Deirdre Bishop
Derek Breese
Enrique Lamas
Jamey Christy
Jennifer Ortman
Jennifer Reichert
John Abowd
Jon Spader
Judy Belton
Karen Battle
Karen Field
Laura Wagoneer
Lindsay Spell
Maryann Chapin
Michael Thieme
Mike Ratcliffe
Pat Cantwell
Roberto Ramirez
Ron Jarmin
Steve Wilson
Stuart Irby
Victoria (Tori) Velkoff
Vincent (Tom) Mule
William (Kin) Koerber

Agenda

1. Update on Status of GQ Work

Meeting Minutes

Identifying Enumerated GQs for Further Investigation

The review team identified census tracts with a decline of 500 or more in GQ population when compared to benchmarks. They also evaluated surrounding census tracts to ensure there was no increase that would offset the observed decline.

This universe of tracts was further evaluated to link GQ units with their facility to evaluate the total facility compared to benchmarks. Internet research was also conducted. Final lists of facilities requiring further investigation were sent to DCMD.

- 20 Military Facilities
  - DCMD has reviewed, utilizing the Joint Services group to make them aware POCs on these bases would be contacted. The contacts are underway.
- 60 Correctional Facilities
  - This includes the state-level prison system (one POC per state). DCMD is reaching out to three states, representing about 30 facilities.
  - DCMD has started calling the local jails (remainder of 60 facilities).
- 20 Nursing Homes
  - DCMD
- 150 Colleges (50 initially sent, followed by 100 additional colleges)

Follow Up for GQs Not Enumerated

GEO has pulled out the list of GQs from the MAF that were not enumerated. Some of the GQs not enumerated that did participate in the GQ advance contact that concluded in February 2020.

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

- Further analysis is being done on the information provided from the advance contact, when available.
- Facilities that were not enumerated and did not have an advance contact number are being contacted.

Contacting Facilities

Calling commenced this morning: NPC is calling colleges. Field is calling nursing facilities.

- Calls are made during business hours.
- Expect colleges to be out starting next week, so staff are working to make all contacts this week.
- Progress will be assessed tonight and again over the weekend to inform decisions about how long to continue this effort.

Addressing Facilities Not Resolved Through Additional Contacts

DSSD is working to develop an imputation procedure.

- DITD will be testing a generic approach for implementing such a procedure.
- A quality metric will be developed, and a benchmark established to determine when to proceed.
- Count imputation for housing units is done on the CUF. The EGG discussed timing for potential count imputation of GQs.
- CES will be asked to scrape information about facility status from the internet (e.g., whether information from the internet indicates a facility is closed) to inform the count imputation model.

A team will be convened by Pat to consider further whether to use count imputation and what that procedure would potentially be. A meeting will be set up as soon as that group is ready to report back to the EGG.

Relevant reports from Census 2000 have been saved to:

[redacted]

Filenames = Census 2000 E.5 R.pdf
Census 2000 TR5.pdf
KK-F-02.pdf

A memo from the 2010 Census is available at:
https://www2.census.gov/programs-surveys/decennial/2010/program-management/5-review/cpex/2010-cpex-243.pdf
(Also saved at [redacted] \2010-cpex-243.pdf)

Jennifer O. will schedule a follow up EGG meeting for 12/15 @ 4pm.

2020 Data Quality Executive Governance Group Meeting
December 15, 2020

EGG Members:

X	Al Fontenot
X	Ben Page
X	Christa Jones
X	Deb Stempowski
X	Enrique Lamas
X	Jamey Christy
X	John Abowd
X	Pat Cantwell
X	Ron Jarmin
X	Tori Velkoff

X = Present

Others:

X	Burton Reist
X	Deirdre Bishop
X	Jennifer Ortman
X	Jennifer Reichert
X	Karen Battle
X	Maryann Chapin
X	Michael Thieme
X	Mike Ratcliffe

X	Andy Keller
X	Anup Mather
X	Juli Zamora
X	Kirk White
X	Nick Pharris-Ciurej
X	Shawn Klimek
X	Sumit Khaneja
X	Tim Kennel
X	Melissa Scopilliti
X	Joseph Staudt

Agenda

1. Update Potential Procedure for GQ Count Imputation

Meeting Minutes

Potential Procedure for GQ Count Imputation

There are a number of GQs (1) classified as occupied with no one enumerated or (2) with counts much lower than benchmarks (including GQ advanced contact). This occurred across all major types of GQs.

An imputation method is being developed for possible application after the creation of DRF2. Data collected from the recent calling operation would merge in with this process.

<u>EGG Questions/Comments</u>

- It may be better to look at GQs at the facility level (rather than unit level). For example, some universities reported all GQ units as one number (facility total rather than subdivided by unit).
  - The DSSD method is focused on GQ unit level. CES is working on the GQ facility level.
- Is there enough information to provide a reasonable expected population estimate for SBEs and transient locations?
  - This is something the teams are giving careful consideration to.
- What proportion of GQ facilities do juvenile facilities account for?
  - Juvenile facilities include correctional facilities, group homes, and treatment centers. Juvenile facilities accounted for about 4% of GQs in 2010.

  - Consideration should be given to which GQ types remain in scope for this work.
- Will most recent ACS data on GQ be used?
  - Yes, information from the ACS is included.
  - The GQ current size is updated each time we get new information for a GQ from either current surveys or a decennial operation. So it can represent 2000, 2010, or an ACS update from 2010-2019.

Meeting Materials

A draft methodology document has been saved to the Data Quality shared drive: [redacted]***\Data Quality EGG 12 15 20 - Group Quarters Imputation Methodology.docx***

Slides are saved at: [redacted]***\20201215 Meeting Materials***

2020 Data Quality Executive Governance Group Meeting
December 23, 2020

EGG Members:

X	Al Fontenot
X	Ben Page
X	Christa Jones
X	Deb Stempowski
X	Enrique Lamas
X	Jamey Christy
X	John Abowd
X	Pat Cantwell
X	Ron Jarmin
X	Tori Velkoff

X = Present

Others:

X	Burton Reist
	Deirdre Bishop
X	Jennifer Ortman
	Jennifer Reichert
X	Karen Battle
X	Maryann Chapin
X	Michael Thieme
X	Mike Ratcliffe

X	Andy Keller
X	Juli Zamora
X	Tim Kennel
X	Shawn Klimek
X	Nick Pharris-Ciurej
X	Kirk White
	Sumit Khaneja
X	Tom Mule
X	Debbie Fenstermaker
X	Mary Frances Zelenak
X	Melissa Scopilliti

Agenda

1. Update Potential Procedure for GQ Count Imputation
2. Update on GQ Overcoverage Work
3. External Engagements (JASON, ASA, Rules of Engagement)

Meeting Minutes

GQ Count Imputation

DSSD has made significant progress to develop models for this work. This includes work to select the base to fit models as well as determining the method for imputing counts once GQs that will be in scope have been identified.

The EGG discussed how to address SBEs in this procedure. The EGG requested additional analysis of shelters be conducted to enable a data-driven decision.

Materials related to this work are available at:
[redacted]

CES has researched external data sources that might be useful to inform the imputation procedure. The IPEDS data seems most promising to assist with imputing counts for dormitories.

GQ Overcoverage

DSSD is evaluating test DRF2 data to identify matches and possible matches. Tom summarized the algorithm that has been developed for this work and the results. The plan is to implement this approach concurrent with completing the analysis.

Materials related to this work are saved at:
[redacted]

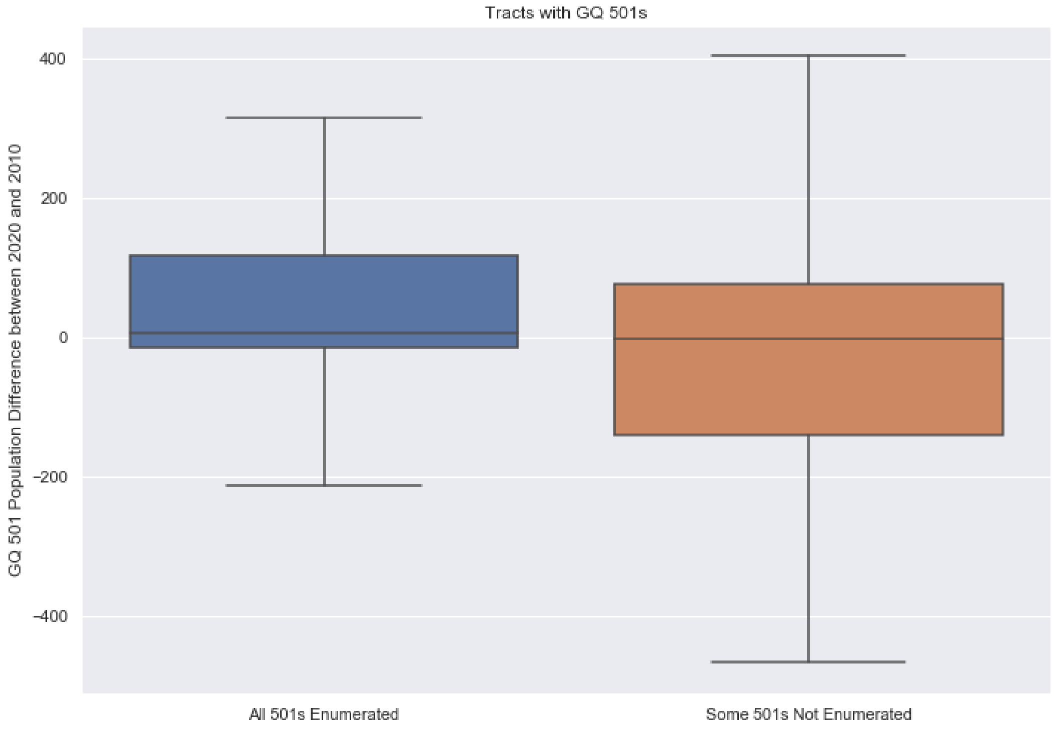
External Engagements

JASON: A series of briefings have been scheduled for the week of January 4, 2021. Census staff will provide background information about the various Census operations with a focus on what is done to assess and ensure quality. Tom Mesenbourg and Herman Habermann have agreed to participate to provide their perspective on quality of census data.

ASA: Conversations are ongoing to establish the scope of this work. Paul Biemer, Bob Fay, and Joe Salvo comprise the team of ASA researchers. There is now a signed scope of work document.

Rules of Engagement: A subset of the EGG met to discuss rules of engagement with external researchers. They reaffirmed that the Census Bureau's standard process will be followed. More information about the guidance that was received is available at:

[redacted] ***2020-12-15 Census Data Quality Rules of Engagement V1.docx***


Tracts with GQ 501s
GQ 501 Population Difference between 2020 and 2010
400
200
0
−200
−400
All 501s Enumerated
Some 501s Not Enumerated

Group Quarters Imputation Methodology

Table of Contents

Table of Tables

Background

There are currently 43,000 MAFIDs classified as occupied group quarters for which we have no reported population count. Errors in individual GQ counts, expecially for larger GQs, are highly visible to the public and could adversely impact the perceived quality of the census. Thus, for large GQs, a count of zero is especially problematic.

A telephone operation is in progress to collect data for some of the larger GQs, including state prisons, local jails, military quarters, nursing homes, and college housing. We will accept all responses from this telephone operation as reported data and will not overwrite these responses with imputed values.

We will impute a GQ population size for the remaining occupied GQs with no reported Census Day population. The occupied group quarters requiring imputation include refusals without any reported people. In addition, group quarters that open on Census Day, but vacant during the GQ Enumeration visit (which started in July 2020) require imputation.

In addition, we will impute a pop size for GQs that have a reported Census Day population count that is much smaller than expected. Our initial proposal is to impute when the Census Day population count is 25% of the GQAC expected count, but research into determining (and refining) this threshould is ongoing.

Imputation Universe

The focus of the GQ Count Imputation is to impute a nonzero count for GQs that are expected to be occupied, but (1) do not have a reported count, or (2) have a reported count that is much smaller than expected. This universe is made up of GQs with a status of Occupied, Vacant During Visit but Open on Census Day, and Refusals. Altogether, we call these GQs unresolved and will impute a count for them. Table 1 shows counts of the GQ universe by GQ status and whether a Census Day population was reported. In Table 1, the GQs with much lower than expected population count are included in the Census Day Pop column. The first three rows represent the occupied GQ universe.

Table 1: GQ Universe

GQ Status	Census Day Pop	No Census Day Pop	Total
Occupied GQ	181,000	17,000	197,000
Open on Census Day, Vacant During Visit	1,900	19,500	21,500
Refusal GQ	1,100	6,700	7,800
Vacant GQ	1,100	29,000	30,500
Delete GQ	450	7,200	7,600
Nonresidential GQ	100	2,400	2,500
Total	185,000	82,000	267,000

Additionally, some of the 185,000 resolved occupied GQs will be treated as unresolved because their census day population is much lower than expected. The goal of the GQ Count Imputation is to determine a population count for all 43,000 unresolved occupied GQs as well as any GQs with a much lower than expected population count. Our current threshold for a “low” population count is < 25% of the GQAC expected count. Table 2 shows the distribution of the resolved and unresolved occupied GQs by GQ status. Of the resolved GQs, 89,0000 had a GQAC expected count and 90,000 did not. The

unresolved GQs include the 43,000 GQs without a reported count as well as 4,500 that had a large discrepancy between the GQAC expected population and the reported pop size.

Table 2: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Status	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	88,500	88,000	3,900	17,000	197,000
Vacant During Visit, Open on Census Day	1,000	550	300	19,500	21,500
Refusal GQ	350	450	300	6,700	7,800
Total	90,000	89,000	4,500	43,000	227,000

Table 3 shows the distribution of the resolved and unresolved occupied GQs by GQ type. Table 10 in the Appendix has a full list of the GQ type codes.

Table 3: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 25% of Expected Pop

GQ Type	Resolved		Unresolved		Total
	No Expected Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	9,900	3,100	300	2,800	16,000
Juvenile Facilities	2,300	3,600	300	1,800	8,000
Nursing Facilities*	6,000	19,000	450	3,200	28,500
Hospitals	750	1,100	100	800	2,800
College Housing*	12,000	17,000	1,400	5,500	36,000
Military*	2,100	900	100	1,900	5,000
Shelters	21,000	3,200	550	8,200	33,000
Group Homes	29,000	32,500	850	9,100	72,000
Other	7,100	8,600	500	9,700	26,000
Total	90,000	89,000	4,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

An alternate definition for a low census day population count would be to use 10% of the GQAC Max Number of People. Table 4 shows counts of the resolved and unresolved cases using this alternate threshold by GQ status. Table 5 shows the same information by GQ type. We will examine using the intersection or union of these conditions as well as setting thresholds at different levels to determine which reported counts require imputation.

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Table 4: Resolved and Unresolved GQ Counts by GQ Status for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Status	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Occupied GQ	67,000	111,000	2,900	17,000	197,000
Vacant During Visit, Open on Census Day	550	1,000	350	19,500	21,500
Refusal GQ	150	650	300	6,700	7,800
Total	68,000	112,000	3,500	43,000	227,000

Note that 2,400 GQs with the Low Census Day Pop based on the Max Pop also have a Low Census Day Pop using the GQAC Expected Population.

Table 5: Resolved and Unresolved GQ Counts by Aggregated GQ Type for Occupied GQs – Low Census Day Pop: < 10 % of GQAC Max Pop

GQ Type	Resolved		Unresolved		Total
	No GQAC Max Pop	Reasonable Census Day Pop	Low Census Day Pop	No Census Day Pop	
Correctional Facilities*	5,600	7,200	400	2,800	16,000
Juvenile Facilities	1,600	4,400	150	1,800	8,000
Nursing Facilities*	4,300	20,500	300	3,200	28,500
Hospitals	550	1,300	90	800	2,800
College Housing*	7,800	21,500	1,200	5,500	36,000
Military*	1,500	1500	90	1,900	5,000
Shelters	17,000	7,300	300	8,200	33,000
Group Homes	24,000	38,500	450	9,100	72,000
Other	5,600	10,000	450	9,700	26,000
Total	68,000	112,000	3,500	43,000	227,000

*denotes GQ Type is included in NPC calling operation

Imputation Methods

Variables

Table 6 shows the variables that are available to impute population counts for the unresolved GQs. Possible sources for data include GQ Advanced Contacts, the current 2020 Decennial Response File 1 (DRF1), the 2010 Census Unedited File (CUF), the American Community Survey, the Master Address File, and Administrative Records. We do not have complete data for any of these auxiliary variables – i.e. each has missing values for at least some of the resolved and unresolved GQs.

Table 6: Auxiliary and Historical Data at the GQ-Level

Variable	Description	Source
GQAC Expected Count	The expected count of people at the Group Quarters on Census Day collected during GQAC.	GQ Advanced Contact
GQAC Max Number of People	The maximum number of people that can live or stay at the Group Quarters at a given time collected during GQAC.	GQ Advanced Contact
Current GQ Size	Number of people at the Group Quarters from the last ACS or Current Surveys visit.	Master Address File / DRF1
Max Number of People	Maximum number of people at the Group Quarters.	Master Address File / DRF1
GQ Type	Based on first digit of GQTYPCUR	DRF1
GQ Status	Occupied GQ; Open on Census Day, Vacant During Visit; Refusal GQ	DRF1
Greek Housing	Indicates whether the GQ name has a Greek letter in the name – signifying a fraternity or sorority house	DRF1
Exists in 2010	Indicates whether the GQ existed on the 2010 CUF	2010 CUF
Occupied HU in 2010	Indicates whether the GQ was an occupied HU on the 2010 CUF	2010 CUF
2010 GQ Count	2010 CUF population count for MAFID	2010 CUF
2010 Occupied HU Count	2010 CUF population count for MAFID	2010 CUF
AR Count	Administrative Records Count (deciles)	AR

Additional sources available for college housing GQs include data collected via web-scraping, data from the Integrated Postsecondary Education Data System (IPEDS) and data from the Common Core. These variables are available at the facility level but not for individual MAFIDs.

We have the 2019 college-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the college level to the GQ data. The main advantage of this variable is that it is available for over 99% of the colleges. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons:

(1) **reference year**—our latest IPEDS data is for reference year 2019;

(2) "**capacity utilization**"—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day;

(3) **scope**---IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Additional facility-level variables may become available as research continues.

Table 7: Facility-Level Data available for College Housing GQs

Variable	Description	Source
Room Cap	If institution provides on-campus housing, the maximum number of students that the institution can provide residential facilities for, whether on or off campus (off-campus dormitory space that is reserved by the institution).	IPEDS

Question: Are there other possible sources or variables (that can be gathered within our timeframe)?

Possible Methods

First, if a pop count is available from the NPC call operation, we will use that pop count as a response and not impute a pop size.

The GQ count imputation will use a combination of the following methods:

1. Ratio Imputation
2. Substitution with Adjusted Residual for College Housing
3. Modeling
4. Median Imputation

Ratio Imputation

For cases where we have an auxiliary count such as an expected GQ pop count as reported in the 2020 Group Quarters Advance Contact (GQAC) operation, we will use ratio imputation. Although the expected GQ count from the GQAC was not reported during the GQ Enumeration (GQE), we believe that such current information (February 2020) may provide a count with less error than other methods. Our research on GQs that reported sufficently during GQE should provide information on this presumption, and on functions of the expected GQ pop count that produce more accurate imputation.

Table 8 shows that 8,600 of the unresolved GQ can be resolved by converting the GQAC expected count to the GQ pop count using the following ratio adjustment.

Table 8: GQ Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

For each GQ type, we will use the ratio of the reported GQ Census Day count to the GQAC expected count to convert the GQAC expected count of the unresolved GQ to a Census Day imputed count. For each GQ type, we will calculate the ratio of the sum of the GQAC Expected Count to the sum of the reported GQ population for the resolved cases. For the unresolved GQs, we will multiply the GQAC expected count by the calculated ratio for that GQ type. For example, for an unresolved College GQ, the following equation would be applied:

$$Imputed\ Population\ Count\ =\ GQAC\ Expected\ Count * \frac{\sum_{GQTYPE=College} Reported\ GQ\ Pop\ Count}{\sum_{GQTYPE=College} GQAC\ Expected\ Count}$$

We will construct ratios in the same manner using the GQAC Max Number of People, Current GQ Size, and Max Number of People variables. We will not use ratio imputation with other prior data, such as the reports from the ACS, IPEDS, or the 2010 Census. Rather, we will use those reported values as covariates to impute a more current pop count. Conversion factors for the four variables under consideration are

shown in Table 9. Tables 12-14 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 9: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0.3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1.017	0.6385
College Housing	0.7818	0.5492	0.9444	0.5535
Military	0.7317	0.2290	0.9492	0.2914
Shelters	0.6261	0.5325	0.6180	0.5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0.3597
All GQs	0.7878	0.5057	0.9217	0.5153

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for GQs for colleges and universities (GQTYPCUR=501).

First, we will adjust the IPEDs room capacity for reference year differences, Greek housing, and for capacity utilization at the college-level, using the Census Day GQ Population, GQAC Max Number of People, and Greek Housing variables.

After adjusting the college-level total room capacity to account reference year and for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Once we calculate the college-level residual, we will then allocate the population counts among the GQs in the college without GQAC Expected Count.

Modeling

A third approach would be to impute the GQ pop counts from a Poisson regression model. The dependent variable will be reported GQ pop count with an offset of the max number of people (because that is filled the most). Independent variables will be selected from Table 6. It is important to note that GQ type will either be a fixed-effect covariate in the models or separate models will be fit by GQ type.

Each model will contain the same set of covariates, with the exception of the college model, which will include additional indicators.

Median Imputation

If sufficient auxiliary data is not available, we will impute the pop size with median population within an imputation cell. This method involves partitioning the GQ universe into imputation cells based on the detailed GQ type and GQ status. Then, we will calculate the median GQ population size and impute the unresolved GQs with the median GQ pop size in the cell.

Question: Are there any other methods we should explore?

Evaluation of Imputed Values

We will evaluate the imputation methods using cross validation. First, we will remove the unresolved GQs from the universe since we don't have a reported GQ pop for them. Second, we will select a stratified systematic sample of occupied GQs. Within each aggregated GQ type, we will select a systematic sample (using max pop count to sort) of 40%. We will call this the training deck. The remaining 60% will be called the validation deck.

We will build and fit our models on the training deck. Then, we will impute the GQ pop size for all GQs in the validation deck. That is, we will attempt to impute the GQ pop size for every GQ in the 60% sample four times (once for each of the four methods). Note that the second method can only be applied to college housing. Then, we will calculate the difference between the reported GQ pop and the imputed GQ pop for each method. We will summarize these differences by computing the minimum of the differences, interquartile range of differences, first quartile of the differences, median of the differences, third quartile of the differences, maximum of the differences, mean of the differences, standard deviation of the differences, and root mean squared error of the differences. We will also produce these metrics for the ratio of the imputed value and the reported value.

Some methods may perform better than others for certain types of units. For example, Poisson regression might perform best when the GQAC expected count is available, but not well when it is missing. Thus, we will calculate the evaluation metrics by GQ types and degrees of missing information to determine the best combination of methods.

Appendix

Table 10: Group Quarter Types

CODE	VALUE
000	Unassigned
010	Campground
020	Recreational Vehicle (RV) Park
030	Marina
040	Hotel or Motel
050	Racetrack
060	Circus or Carnival
090	Other Transitory Location
101	Federal Detention Centers
102	Federal Prisons
103	State Prisons
104	Local Jails and Other Municipal Confinement Facilities
105	Correctional Residential Facilities
106	Military Disciplinary Barracks and Jails
201	Group Homes for Juveniles (non-correctional)
202	Residential Treatment Centers for Juveniles (non-correctional)
203	Correctional Facilities Intended for Juveniles
301	Nursing Facilities/Skilled Nursing Facilities
401	Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals
402	Hospitals with Patients Who Have No Usual Home Elsewhere
403	In-Patient Hospice Facilities
404	Military Treatment Facilities with Assigned Patients
405	Residential Schools for People with Disabilities
501	College/University Student Housing (College/University owned/leased/managed)
502	College/University Student Housing (Privately owned/leased/managed)
601	Military Quarters
602	Military Ships
701	Emergency and Transitional Shelters (With Sleeping Facilities) for People Experiencing Homelessness
702	Soup Kitchens
704	Regularly Scheduled Mobile Food Vans
706	Targeted Non-Sheltered Outdoor Locations
801	Group Homes Intended for Adults
802	Residential Treatment Centers for Adults
900	Maritime/Merchant Vessels
901	Workers' Group Living Quarters and Job Corps Centers
903	Living Quarters for Victims of Natural Disaster
904	Other Noninstitutional Group Quarters
999	Unknown Group Quarters Type

Table 11: GQAC Expected Count by Imputation Status

GQAC Expected Count	Resolved	Unresolved	Total
Populated	93,500	8,600	102,000
Not Populated	90,000	34,500	125,000
Total	184,000	43,000	227,000

Table 12: GQAC Max Number of People by Imputation Status

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Table 13: Current GQ Size by Imputation Status

Current GQ Size	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

Table 14: Max Number of People by Imputation Status

Max Number of People	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

Update on Group Quarters Count Imputation

Group Quarters Count Imputation Team

Meeting of Data Quality EGG

12/15/20

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU

DRB Approval Number: CBDRB-FY21-DSEP-002

The Problem

- For Group Quarters (GQs) in our data processing system
  - More than 40,000 GQ MAFIDs are classified as occupied, but with no reported population count
  - Others with a pop count much smaller than expected, perhaps 3,000 to 7,000

- Some details
  - Occurs across all major types of GQs
  - Includes refusals; occupied; and open on Census Day, but vacant during visit
  - Assumption: we'll retain (not impute for) responses from call operation

DRB Approval Number: CBDRB-FY21-DSEP-002

Addressing the Problem (1 of 3)

- Team cuts across several directorates and divisions
  - Staff from DITD, CES, DSSD

- Processing GQ-level files
  - Will produce a file with GQ MAFID and imputed pop count—number of records to be created following DRF2; leads into CUF processing
  - Will combine this file with others derived from fixing other GQ problems
    - results from recent calling operation
    - data collected from 35,000 ICQs, paper listings

DRB Approval Number: CBDRB-FY21-DSEP-002

Addressing the Problem (2 of 3)

- Information available
  - For some GQs, information from GQ Advance Contact: expected, max count
  - Other internal data, e.g., data on these GQs from ACS or 2010 Census
  - Data from official sources, Integrated Postsecondary Education Data System
  - Data from internet, including web scraping
- Methodology (more later)
  - Investigating several approaches, all within different major types of GQ
  - Will develop models on good data, evaluate them on other good data
  - Will examine the models, apply on missing data, assess results, narrow focus

DRB Approval Number: CBDRB-FY21-DSEP-002

Addressing the Problem (3 of 3)

- Timeline, very aggressive
  - by 12/23/20, complete research, testing, validation, selection of models
  - by 12/24/20, run (execute) models for production
  - by 12/29/20, complete review by SMEs in POP and DSSD

- Questions and considerations
  - For which types of GQ do we impute?
  - For GQs with questionable response, threshold for imputing? how conservative?
  - Procedures: hierarchical? how intricate?
  - For later: approach for imputing characteristics on Census Edited File

DRB Approval Number: CBDRB-FY21-DSEP-002

DRB Approval Number: CBDRB-FY21-DSEP-002

Reasonableness Reviews and Data Quality: Briefing Plans + Initial Findings

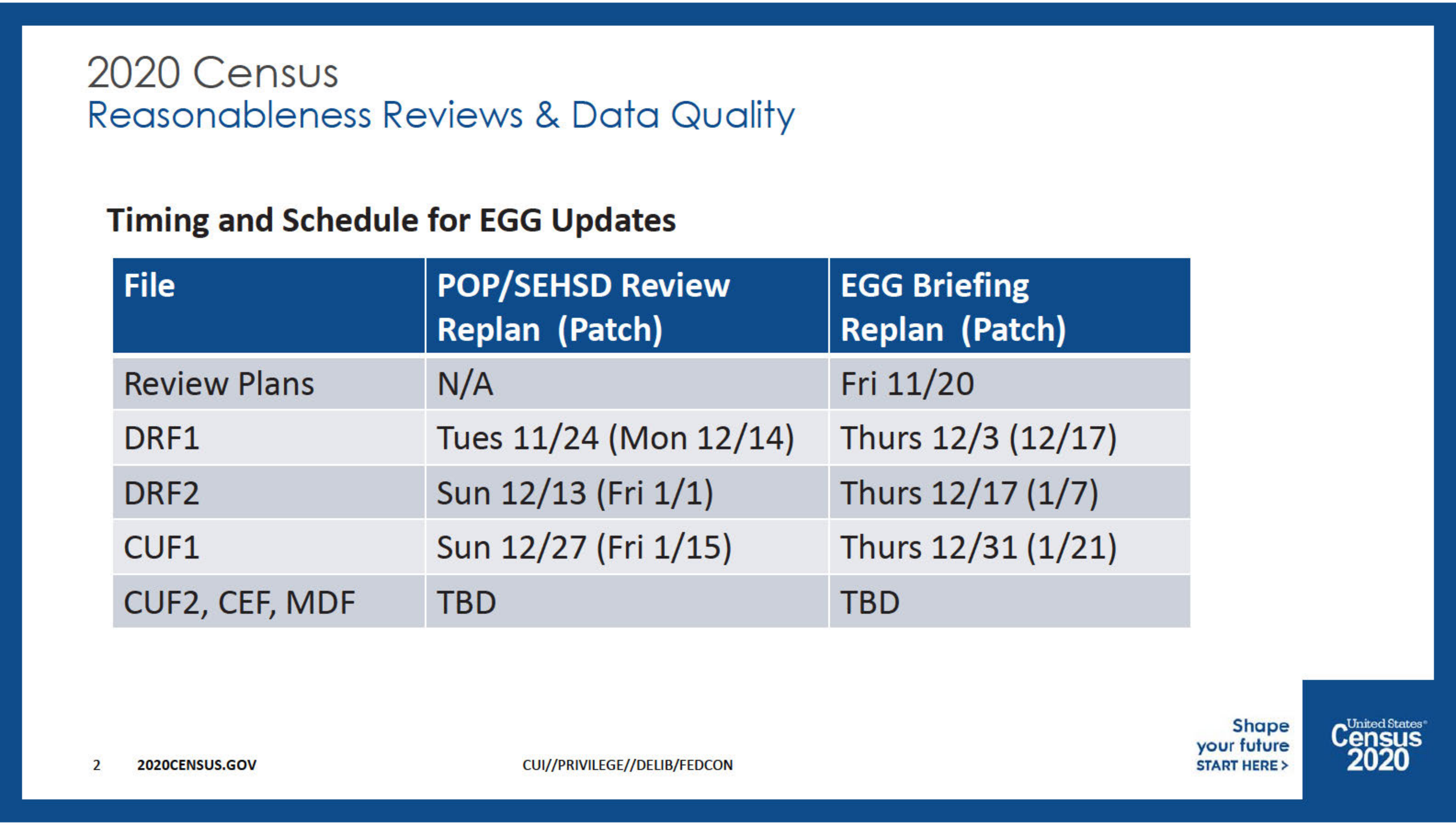
Jonathan Spader, SEHSD
Christine Borman, POP

November 20, 2020

Shape
your future
START HERE >

United States®
Census
2020

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



2020 Census
Reasonableness Reviews & Data Quality

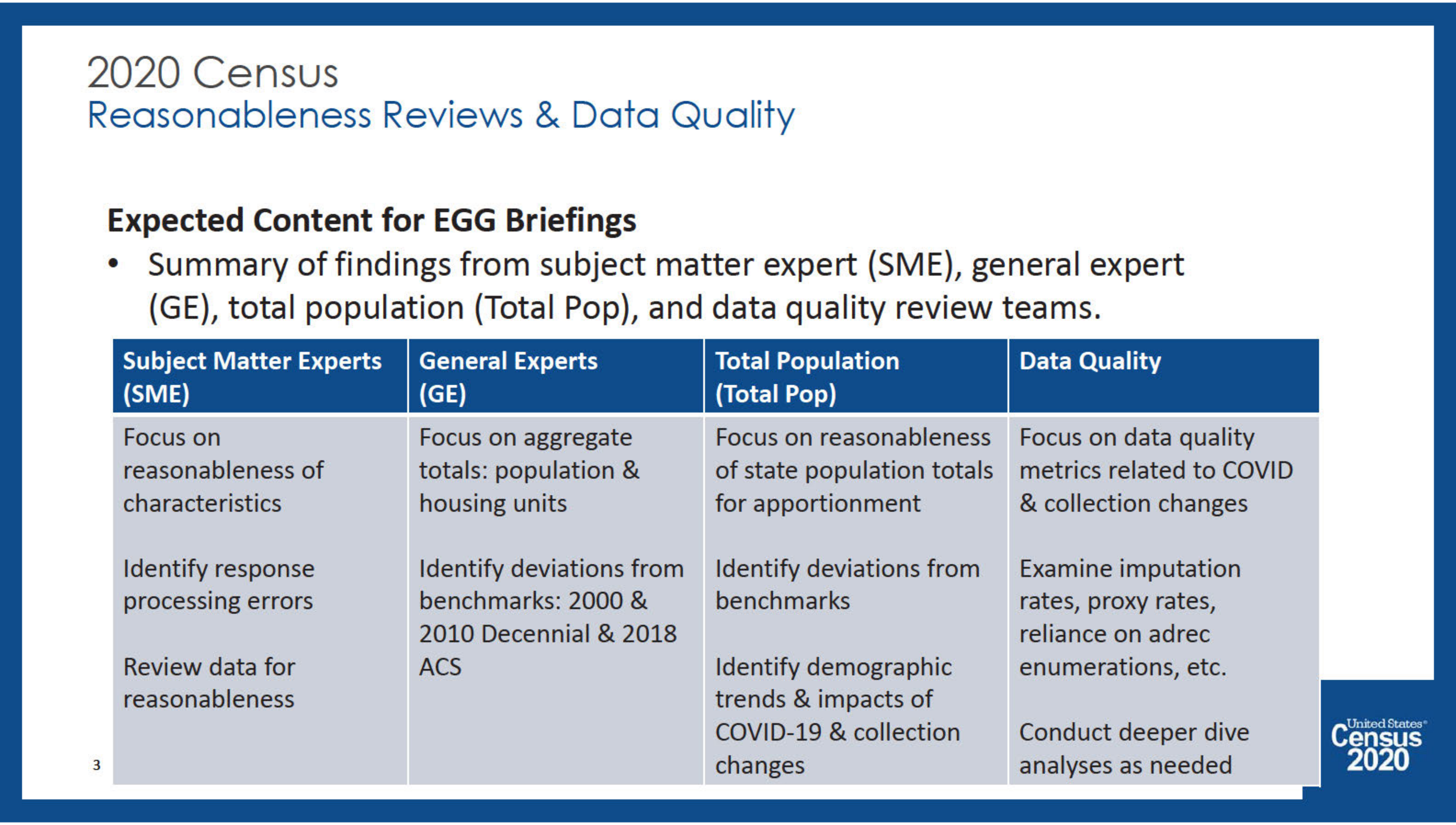
Timing and Schedule for EGG Updates

File	POP/SEHSD Review Replan (Patch)	EGG Briefing Replan (Patch)
Review Plans	N/A	Fri 11/20
DRF1	Tues 11/24 (Mon 12/14)	Thurs 12/3 (12/17)
DRF2	Sun 12/13 (Fri 1/1)	Thurs 12/17 (1/7)
CUF1	Sun 12/27 (Fri 1/15)	Thurs 12/31 (1/21)
CUF2, CEF, MDF	TBD	TBD

Shape your future START HERE >

United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



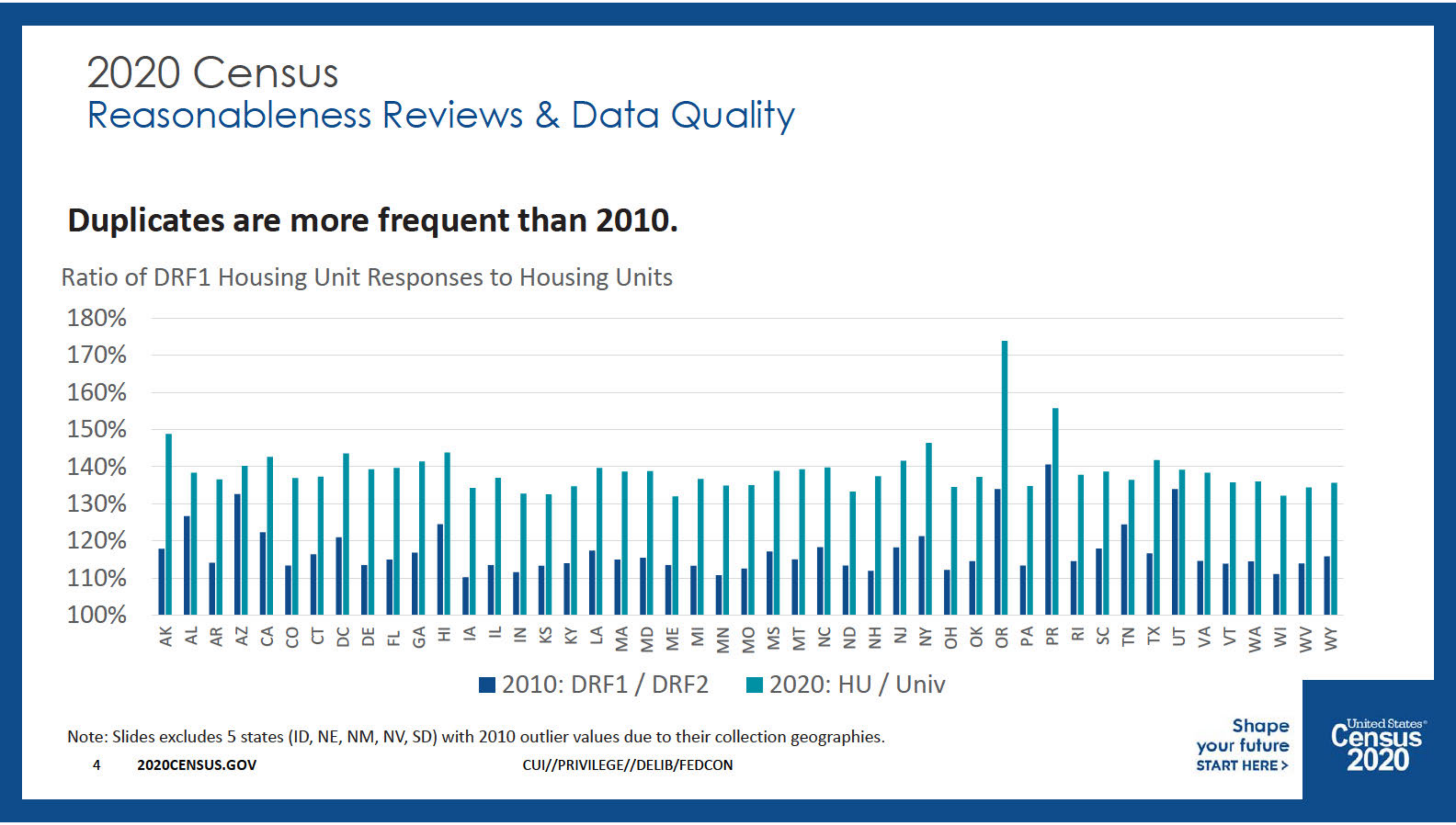
2020 Census
Reasonableness Reviews & Data Quality

Expected Content for EGG Briefings

- Summary of findings from subject matter expert (SME), general expert (GE), total population (Total Pop), and data quality review teams.

Subject Matter Experts (SME)	General Experts (GE)	Total Population (Total Pop)	Data Quality
Focus on reasonableness of characteristics Identify response processing errors Review data for reasonableness	Focus on aggregate totals: population & housing units Identify deviations from benchmarks: 2000 & 2010 Decennial & 2018 ACS	Focus on reasonableness of state population totals for apportionment Identify deviations from benchmarks Identify demographic trends & impacts of COVID-19 & collection changes	Focus on data quality metrics related to COVID & collection changes Examine imputation rates, proxy rates, reliance on adrec enumerations, etc. Conduct deeper dive analyses as needed

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



2020 Census
Reasonableness Reviews & Data Quality

Duplicates are more frequent than 2010.

Ratio of DRF1 Housing Unit Responses to Housing Units

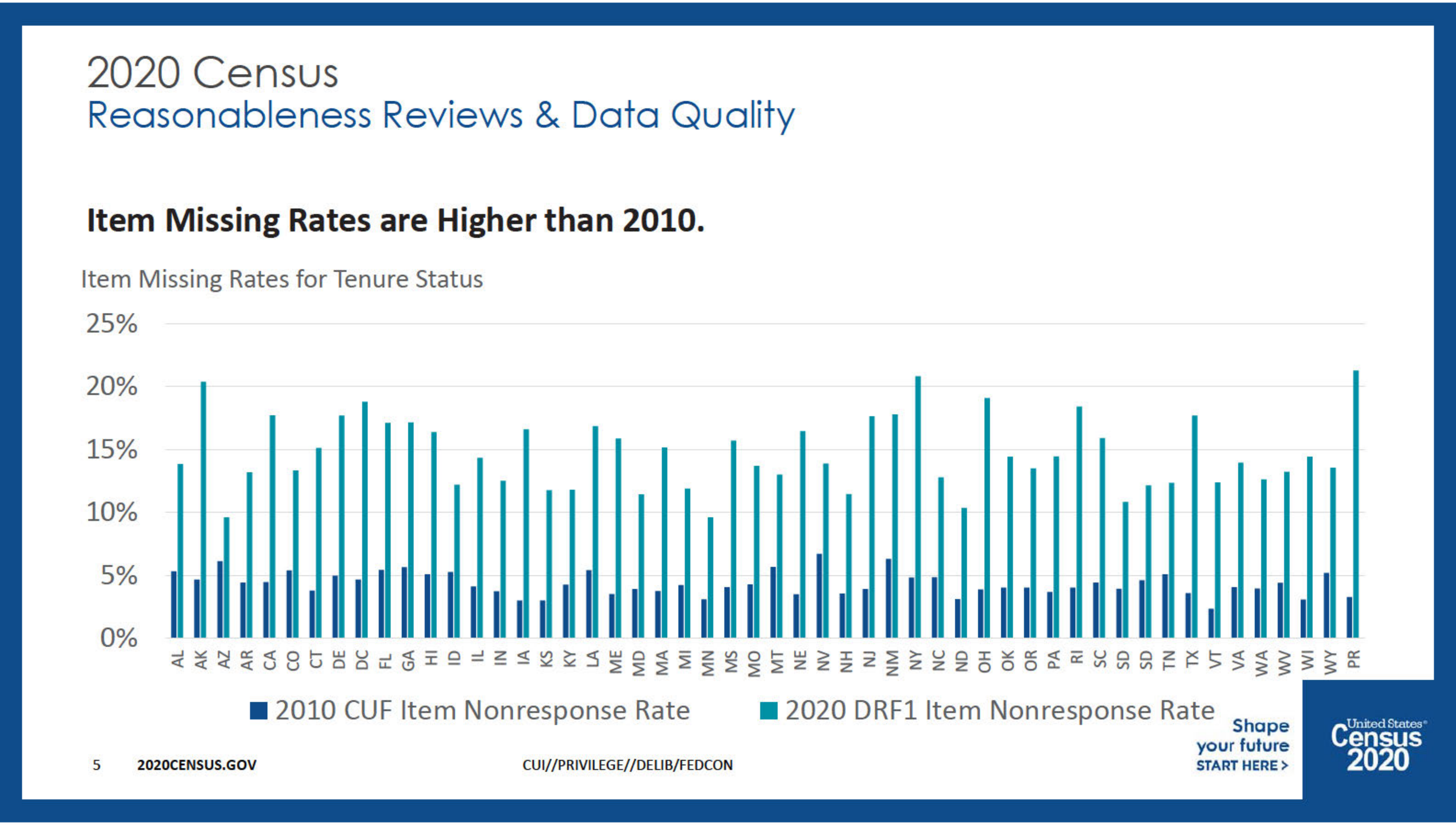
Note: Slides excludes 5 states (ID, NE, NM, NV, SD) with 2010 outlier values due to their collection geographies.

CUI//PRIVILEGE//DELIB/FEDCON

Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



2020 Census
Reasonableness Reviews & Data Quality

Item Missing Rates are Higher than 2010.

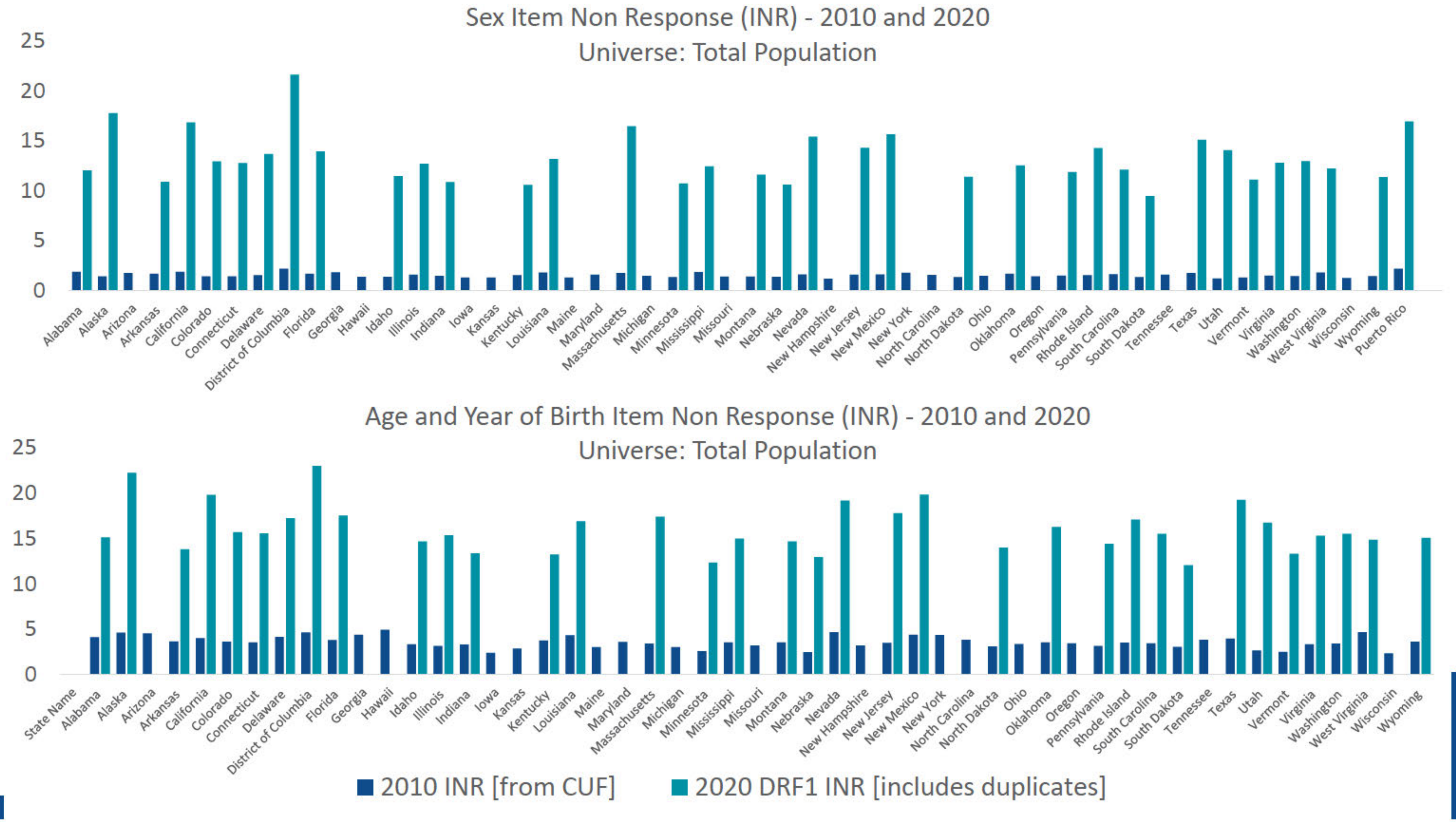
Item Missing Rates for Tenure Status

Shape your future START HERE >

United States Census 2020

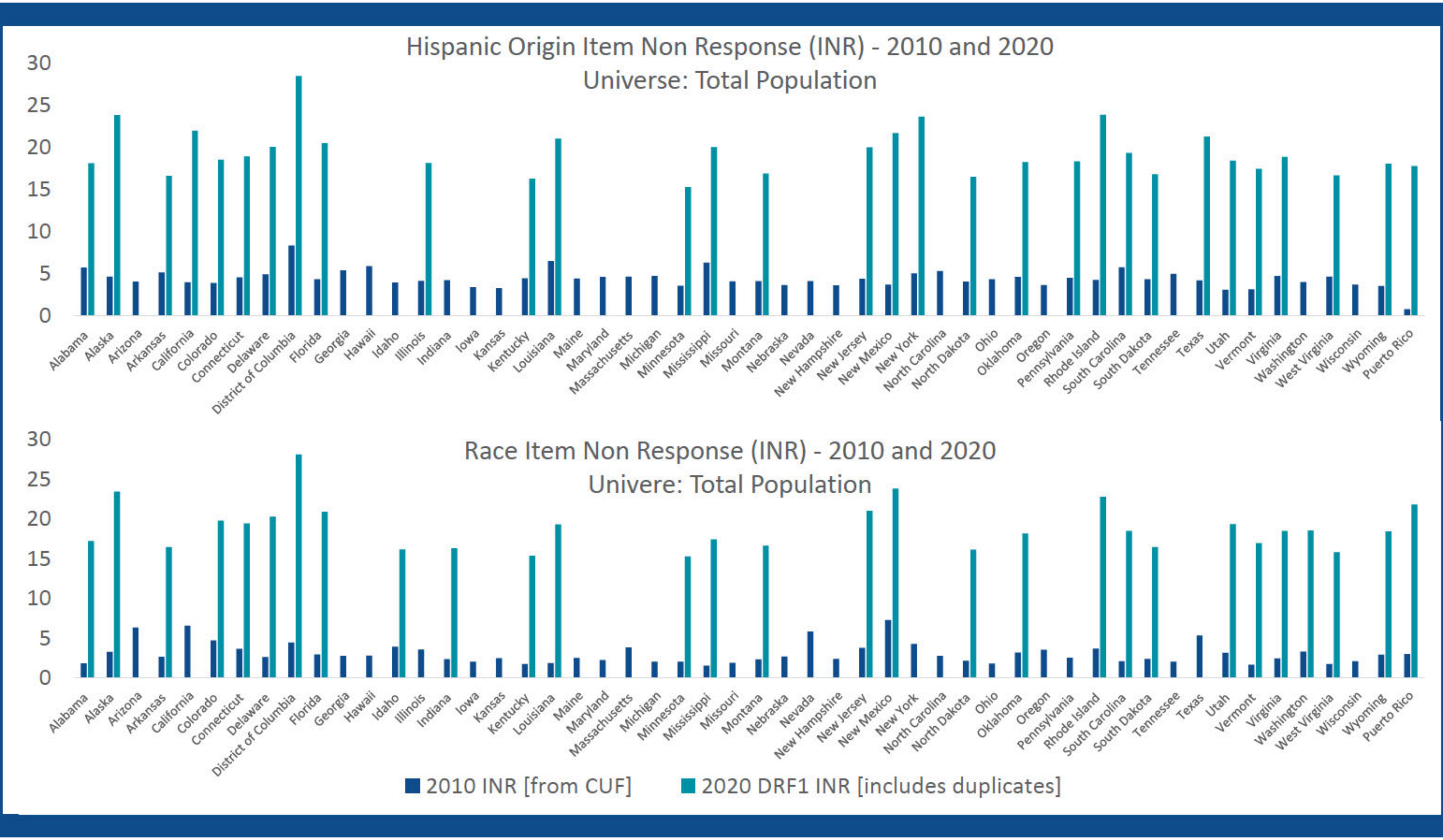
CUI//PRIVILEGE//DELIB/FEDCON

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.


Sex Item Non Response (INR) - 2010 and 2020
Universe: Total Population

Age and Year of Birth Item Non Response (INR) - 2010 and 2020
Universe: Total Population

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



Hispanic Origin Item Non Response (INR) - 2010 and 2020
Universe: Total Population

Race Item Non Response (INR) - 2010 and 2020
Univere: Total Population

■ 2010 INR [from CUF] ■ 2020 DRF1 INR [includes duplicates]

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

District of Columbia – DRF1

Data Collection Mode	N	Missing Sex	Missing Age and Year of Birth	Missing Hispanic Origin	Missing Race
2010 Census Total Pop INR from the 2010 CUF	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
2020 DRF1 Total Population	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Internet Self Response (ISR)	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Paper Self Response	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
NRFU Production	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Census Questionnaire Assistance (CQA)	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Coverage Improvement (CI)	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
GQ eResponse	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
GQ Facility Self-enumeration	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
GQ Paper Listing	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
NRFU Administrative Records Enumeration	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Note: Responses from NRFU Reinterview and NRFU Response Validation (SRQA) are expected to have high missing rates and are therefore excluded from the mode lines.
GQ and HU dummy records created in post-processing are excluded from this table; they have 100% missing rates.

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

South Carolina - Hispanic Origin and Race Reasonableness

[redacted]

Not Hispanic Hispanic No Response

■ 2010 CUF ■ 2019 Estimates ■ 2020 DRF1

White alone Black alone AIAN alone Asian alone NHPI alone SOR alone Two or More Races No Response

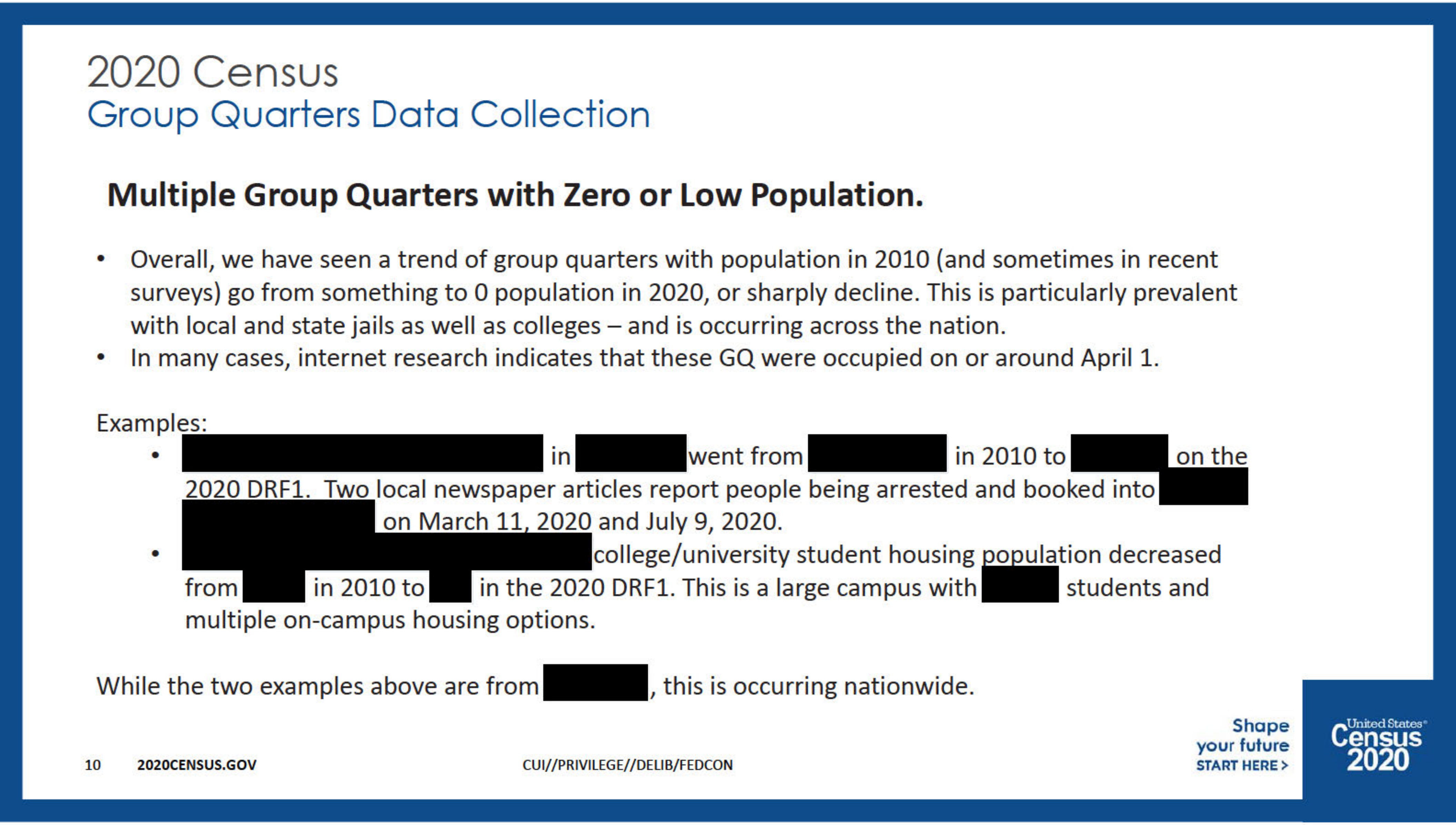
■ 2010 CUF ■ 2019 ACS ■ 2020 DRF1

Note: Hispanic origin and race groups determined by checkbox responses only.

Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



2020 Census
Group Quarters Data Collection

Multiple Group Quarters with Zero or Low Population.

- Overall, we have seen a trend of group quarters with population in 2010 (and sometimes in recent surveys) go from something to 0 population in 2020, or sharply decline. This is particularly prevalent with local and state jails as well as colleges – and is occurring across the nation.
- In many cases, internet research indicates that these GQ were occupied on or around April 1.

Examples:

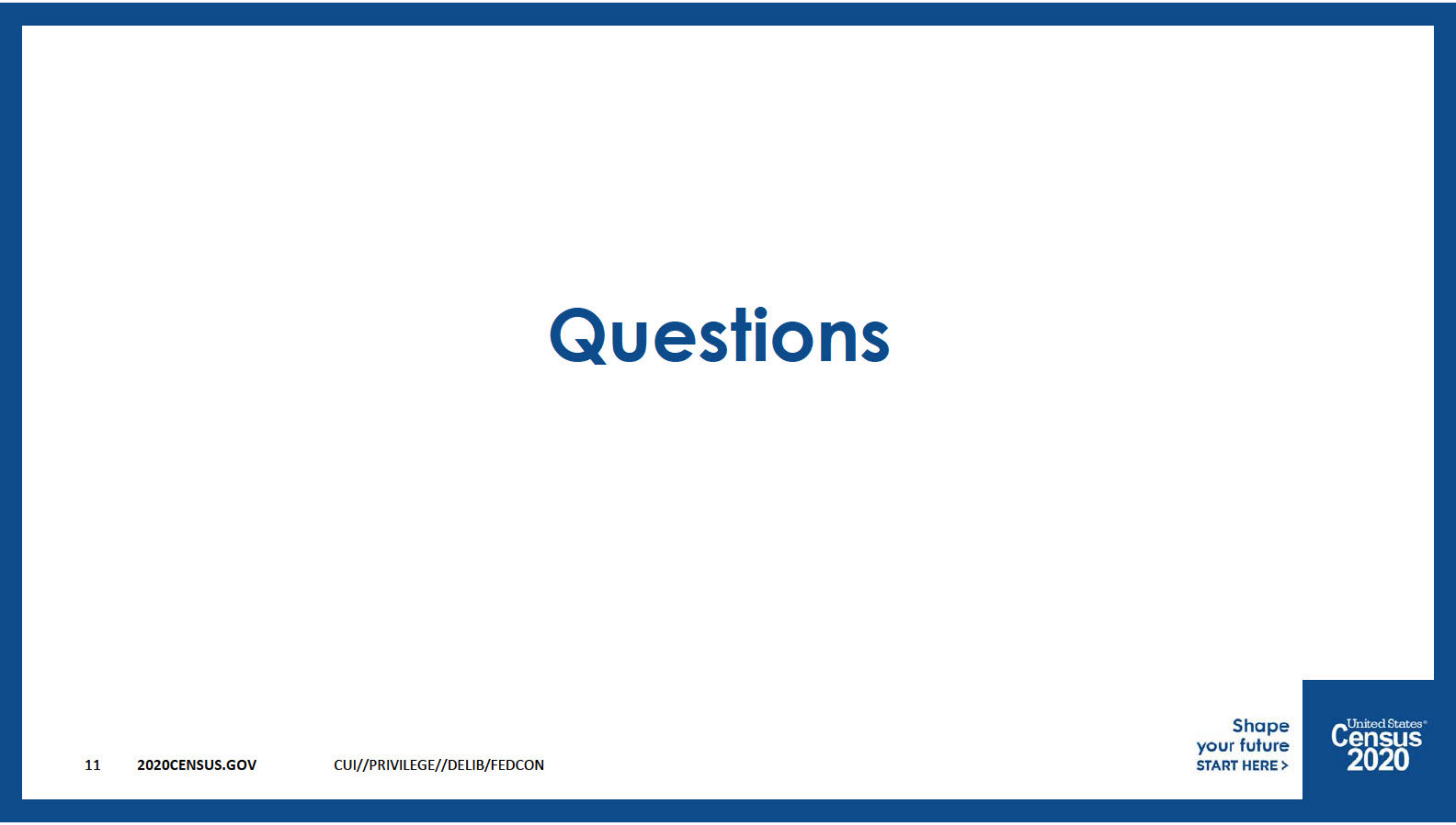
- [redacted] in [redacted] went from [redacted] in 2010 to [redacted] on the 2020 DRF1. Two local newspaper articles report people being arrested and booked into [redacted] on March 11, 2020 and July 9, 2020.
- [redacted] college/university student housing population decreased from [redacted] in 2010 to [redacted] in the 2020 DRF1. This is a large campus with [redacted] students and multiple on-campus housing options.

While the two examples above are from [redacted], this is occurring nationwide.

Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



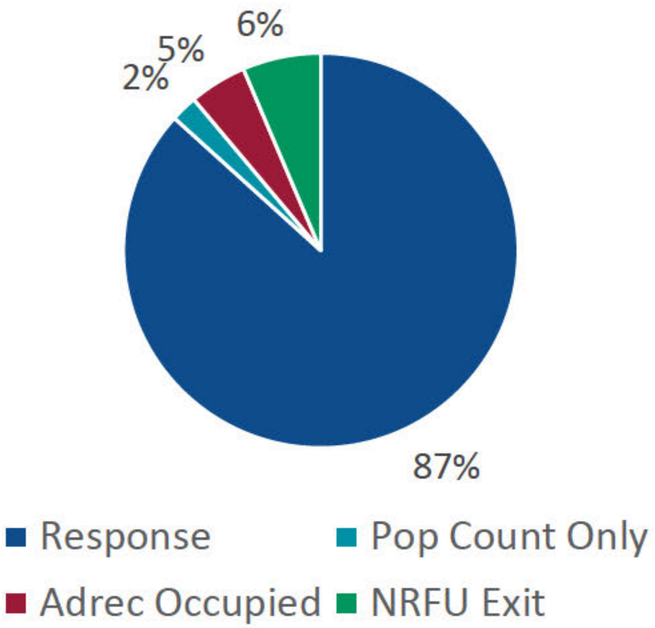
Questions

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

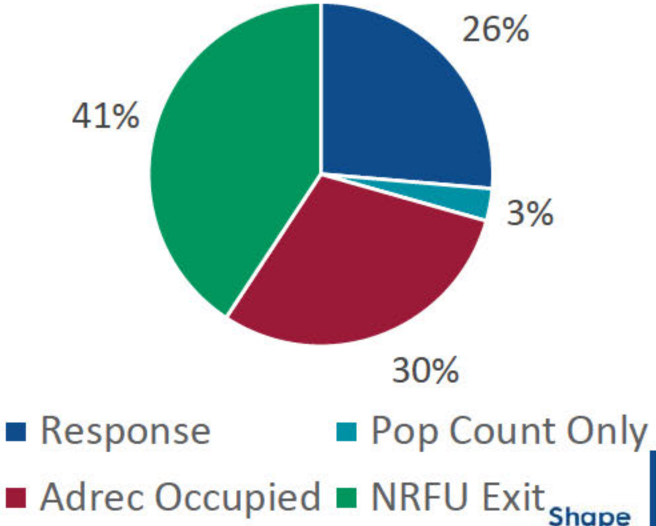
2020 Census
Reasonableness Reviews & Data Quality

71% of Missing Tenure Values are due to Adrec Occupied & NRFU Exits.

All Occupied Housing Unit Responses



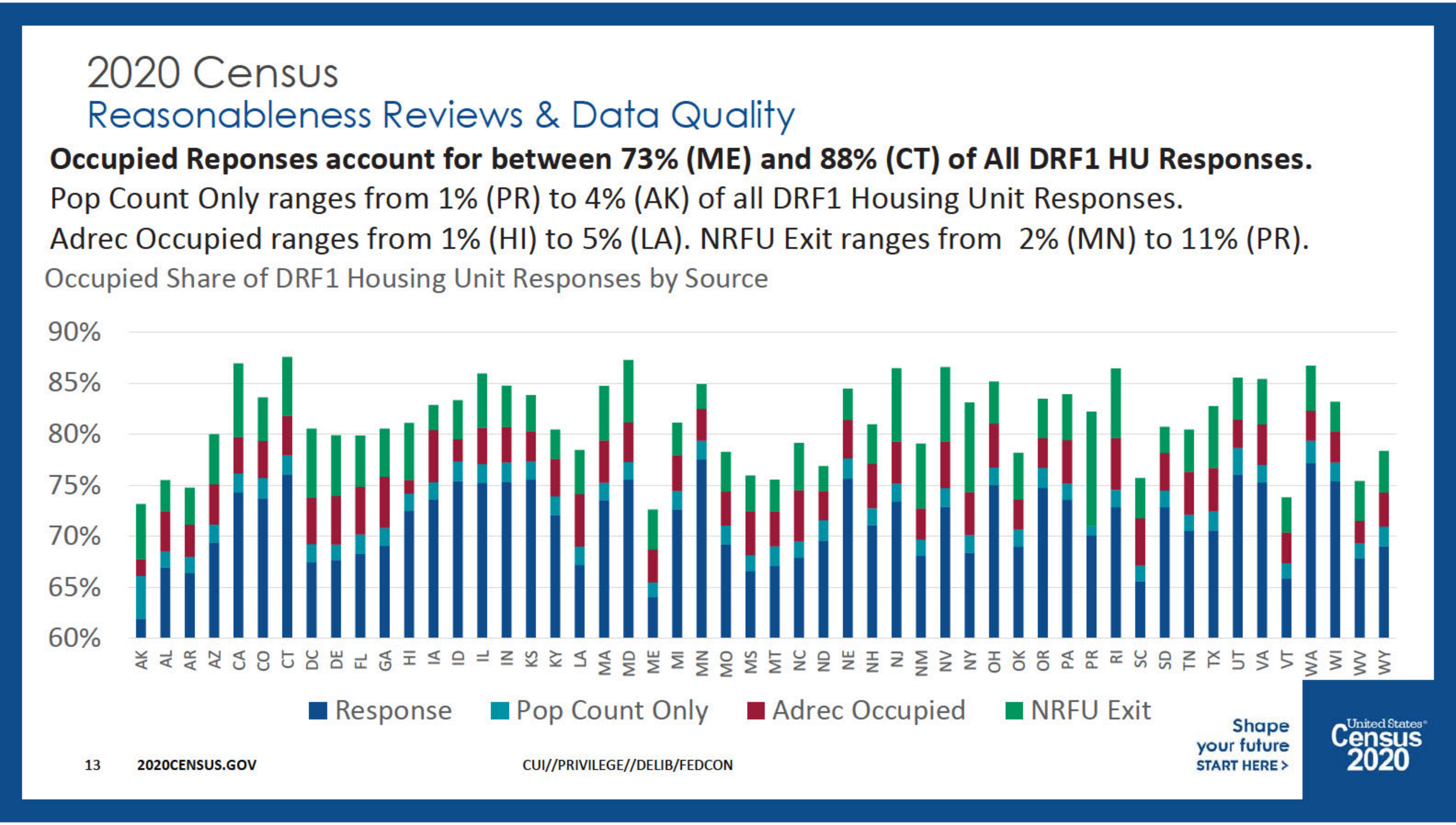
Occupied Housing Unit Responses with Missing Tenure



Shape your future START HERE >

United States® Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



2020 Census

Reasonableness Reviews & Data Quality

Occupied Reponses account for between 73% (ME) and 88% (CT) of All DRF1 HU Responses.

Pop Count Only ranges from 1% (PR) to 4% (AK) of all DRF1 Housing Unit Responses.

Adrec Occupied ranges from 1% (HI) to 5% (LA). NRFU Exit ranges from 2% (MN) to 11% (PR).

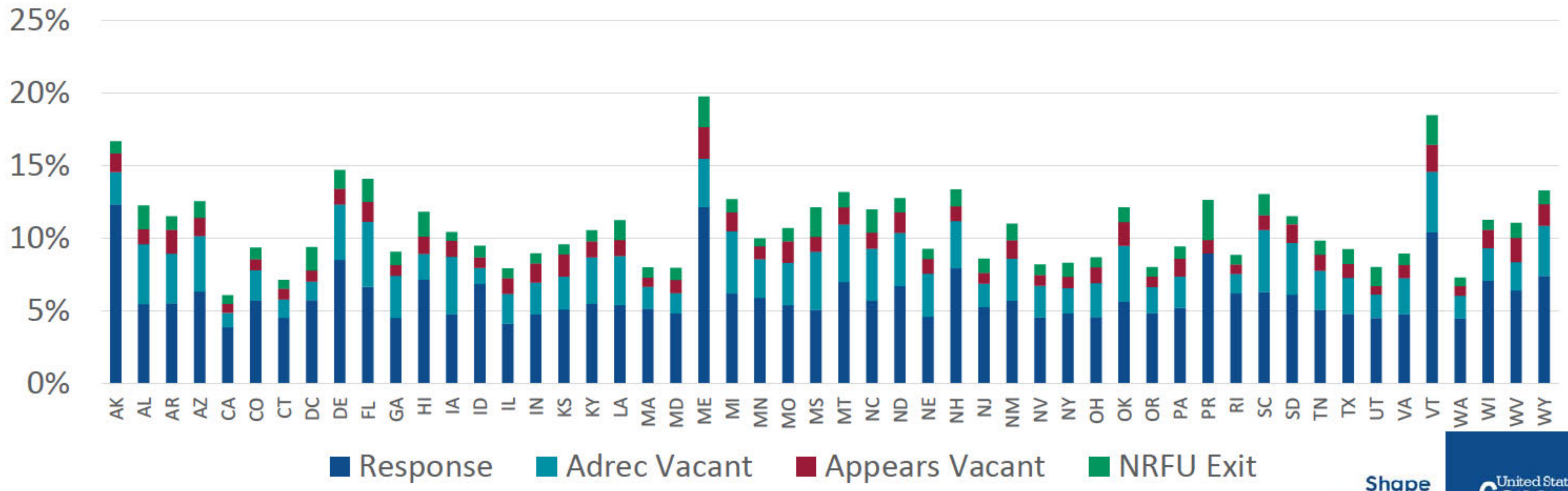
Occupied Share of DRF1 Housing Unit Responses by Source

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

2020 Census
Reasonableness Reviews & Data Quality

Vacant Reponses account for between 6% (CA) and 20% (ME) of DRF1 Responses.

Vacant Share of DRF1 Housing Unit Responses by Source



Shape your future START HERE >

United States Census 2020

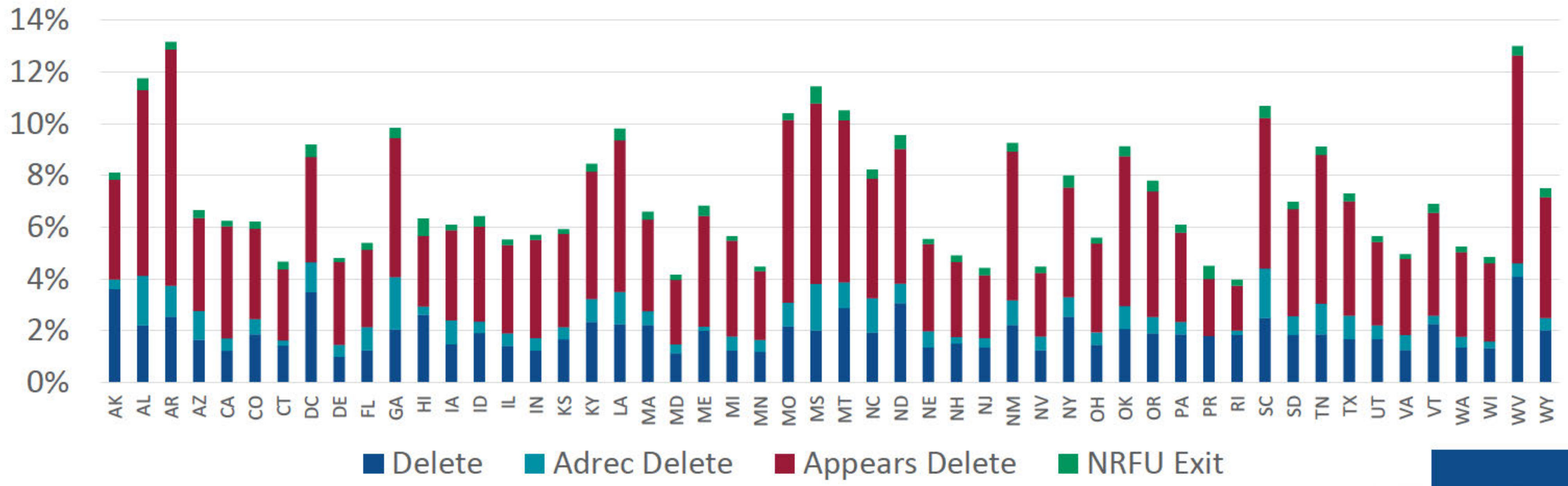
14 2020CENSUS.GOV CUI//PRIVILEGE//DELIB/FEDCON

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

2020 Census
Reasonableness Reviews & Data Quality

Delete Reponses account for between 4% (RI) and 13% (AR) of DRF1 Responses.

Delete Share of DRF1 Housing Unit Responses by Source



CUI//PRIVILEGE//DELIB/FEDCON

Shape your future START HERE >

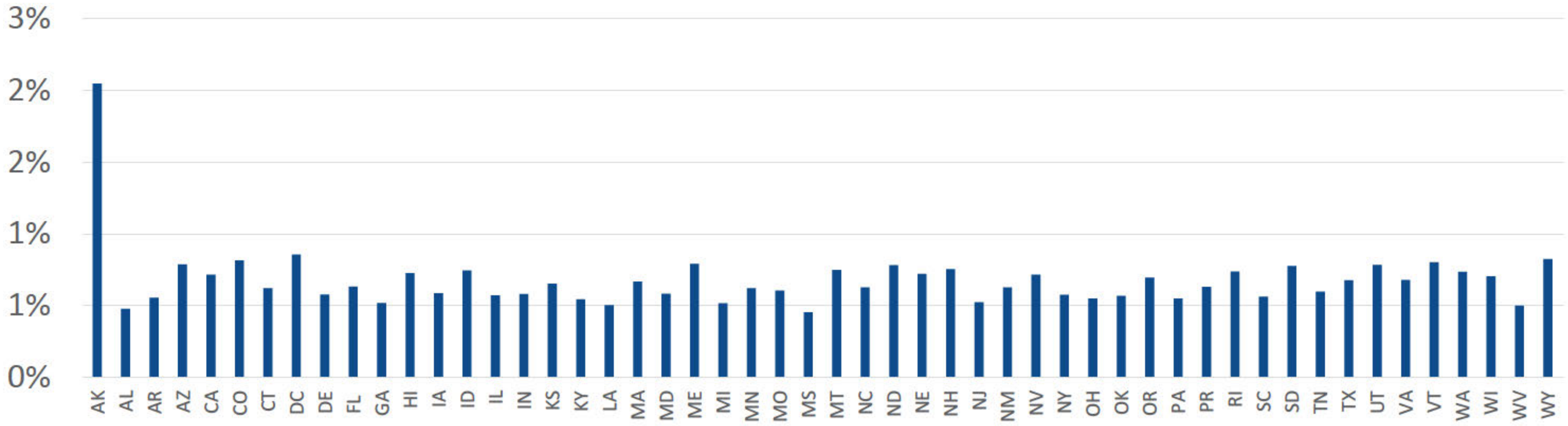
United States Census 2020

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

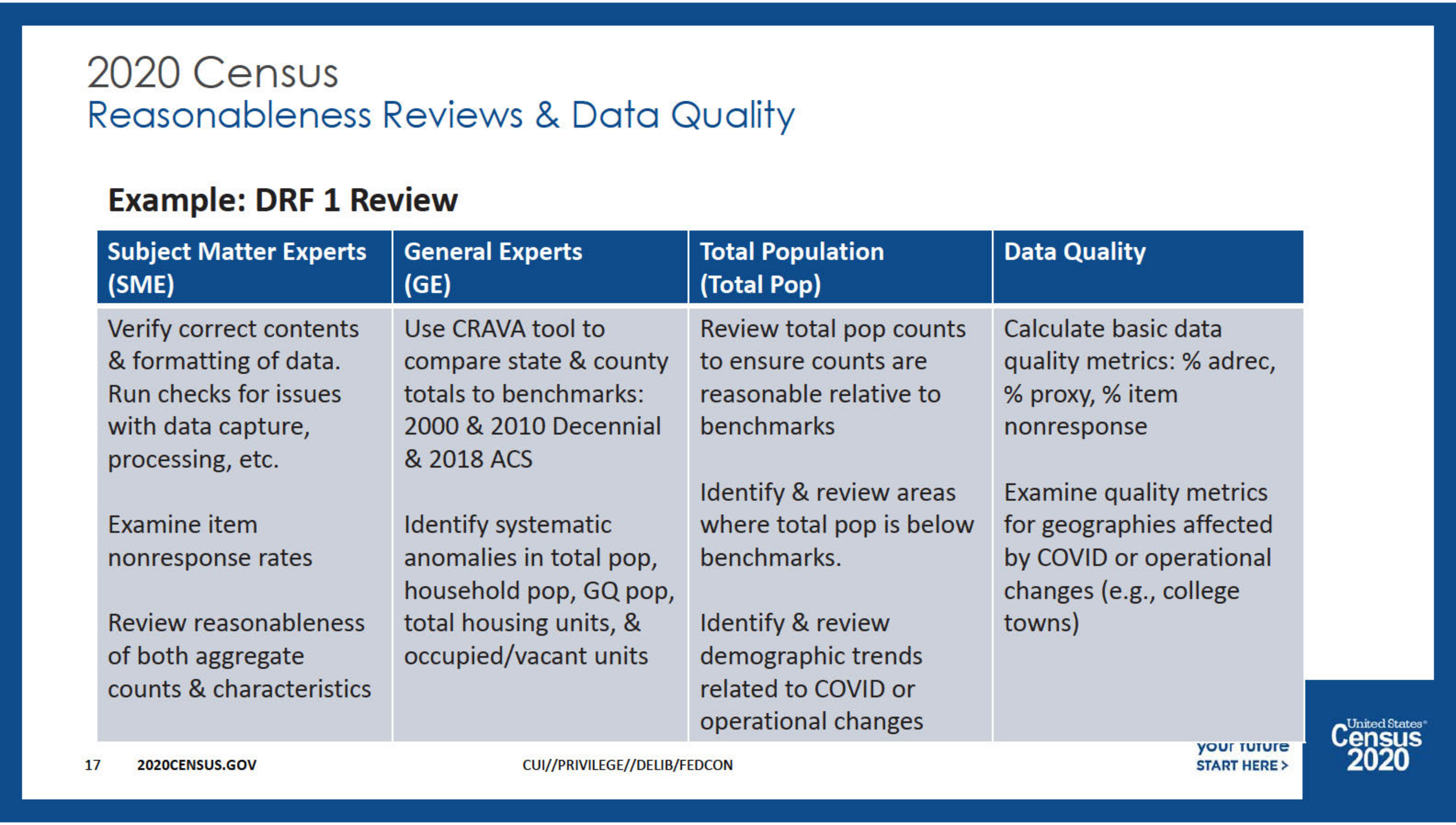
2020 Census
Reasonableness Reviews & Data Quality

Unresolved Reponses account for between 0.5% (MS) and 2.0% (AK) of DRF1 Responses.

Unresolved Share of DRF1 Housing Unit Responses



DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.



2020 Census
Reasonableness Reviews & Data Quality

Example: DRF 1 Review

Subject Matter Experts (SME)	General Experts (GE)	Total Population (Total Pop)	Data Quality
Verify correct contents & formatting of data. Run checks for issues with data capture, processing, etc. Examine item nonresponse rates Review reasonableness of both aggregate counts & characteristics	Use CRAVA tool to compare state & county totals to benchmarks: 2000 & 2010 Decennial & 2018 ACS Identify systematic anomalies in total pop, household pop, GQ pop, total housing units, & occupied/vacant units	Review total pop counts to ensure counts are reasonable relative to benchmarks Identify & review areas where total pop is below benchmarks. Identify & review demographic trends related to COVID or operational changes	Calculate basic data quality metrics: % adrec, % proxy, % item nonresponse Examine quality metrics for geographies affected by COVID or operational changes (e.g., college towns)

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

DRF1 Review Update:
Total Population and Group Quarters

Marc Perry, POP
Christine Borman, POP

December 3, 2020

DRB Approval Number: CBDRB-FY21-DSEP-002

[redacted]

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002

[redacted]

nited States®
ensus
2020

Group Quarters (GQ) Review

We focus on GQ data now because we anticipate considerable unduplication of the household population in DRF2. The benefits of looking at GQ data now:

1. **Relatively stable populations** over time for many large and well known GQ facilities.
2. Estimates from **Group Quarters Advanced Contact (GQAC)** prior to data collection and other benchmarks give us a good picture of what the population for that GQ could be.
3. **Unduplication during DRF2 is done within GQ unit – not across GQ units** in an overall GQ facility. Three Christine Bormans within a college dorm (i.e. a GQ unit) would be unduplicated whereas three Christine Bormans in the entire college with multiple dorms (i.e. the GQ facility) would not be unduplicated. **So we know when we see an issue across units within a facility, it is likely to remain an issue after further processing.**

One challenge we do have with GQs is that the **units within a GQ facility are not clearly linked.** We could try to look at GQ facility names, but many we looked at contained typos in the facility name that made them difficult to link.

GQ populations will still change during DRF2 processing, but so far, we've seen some of the following issues…

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002

Pre-decisional - Internal Use Only - Not for Public Distribution - Disclosure Prohibited T-13 U.S. Code

[redacted]

Pre-decisional - Internal Use Only - Not for Public Distribution - Disclosure Prohibited T-13 U.S. Code

Topics for future presentations

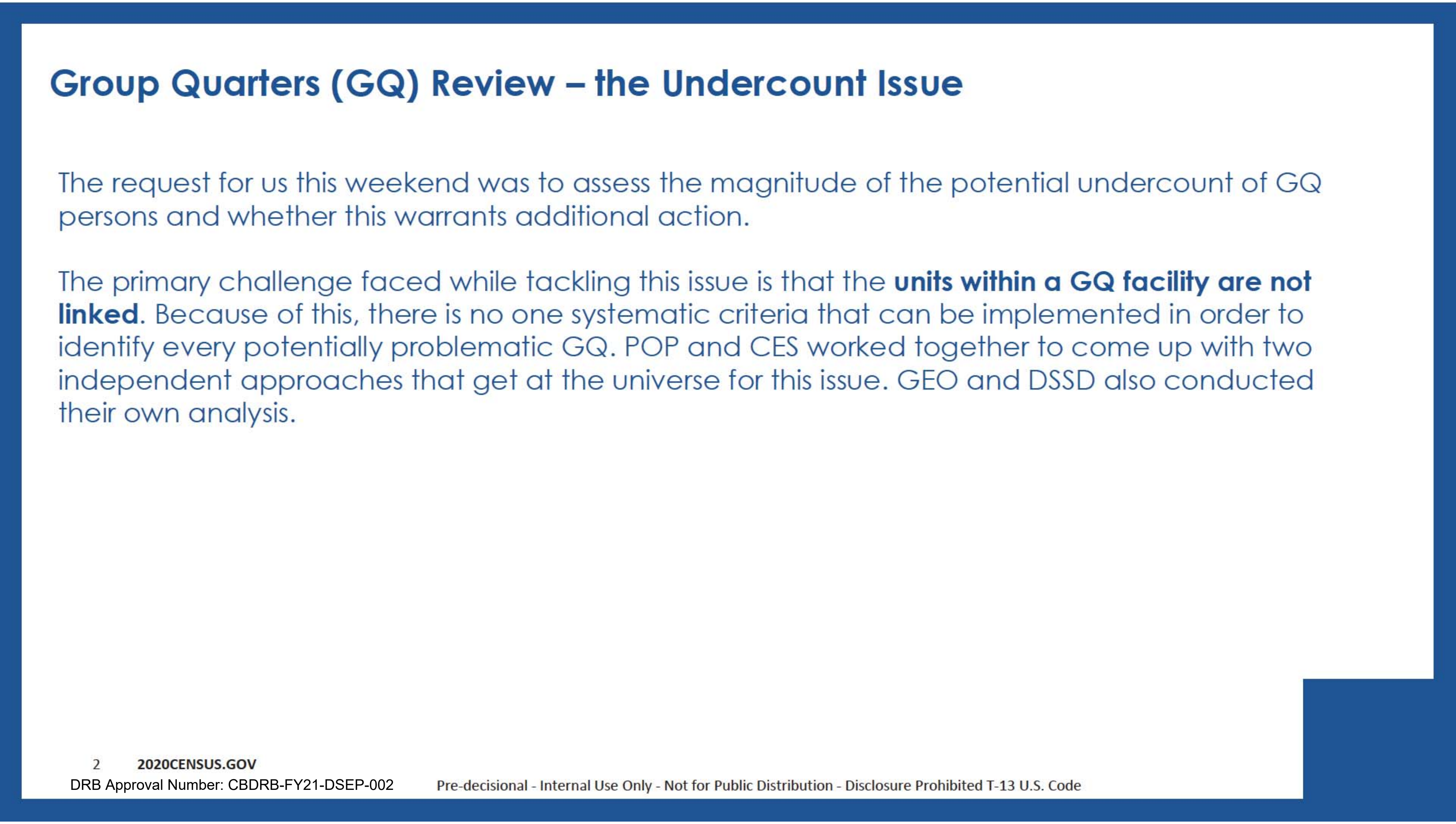
- Duplicate analysis
- Item non response missing rates
- Off-campus college analysis and other geographies

Any other questions or areas the EGG would like POP and SEHSD to cover?

Group Quarters Undercount Issue Update

Population Division
December 6, 2020

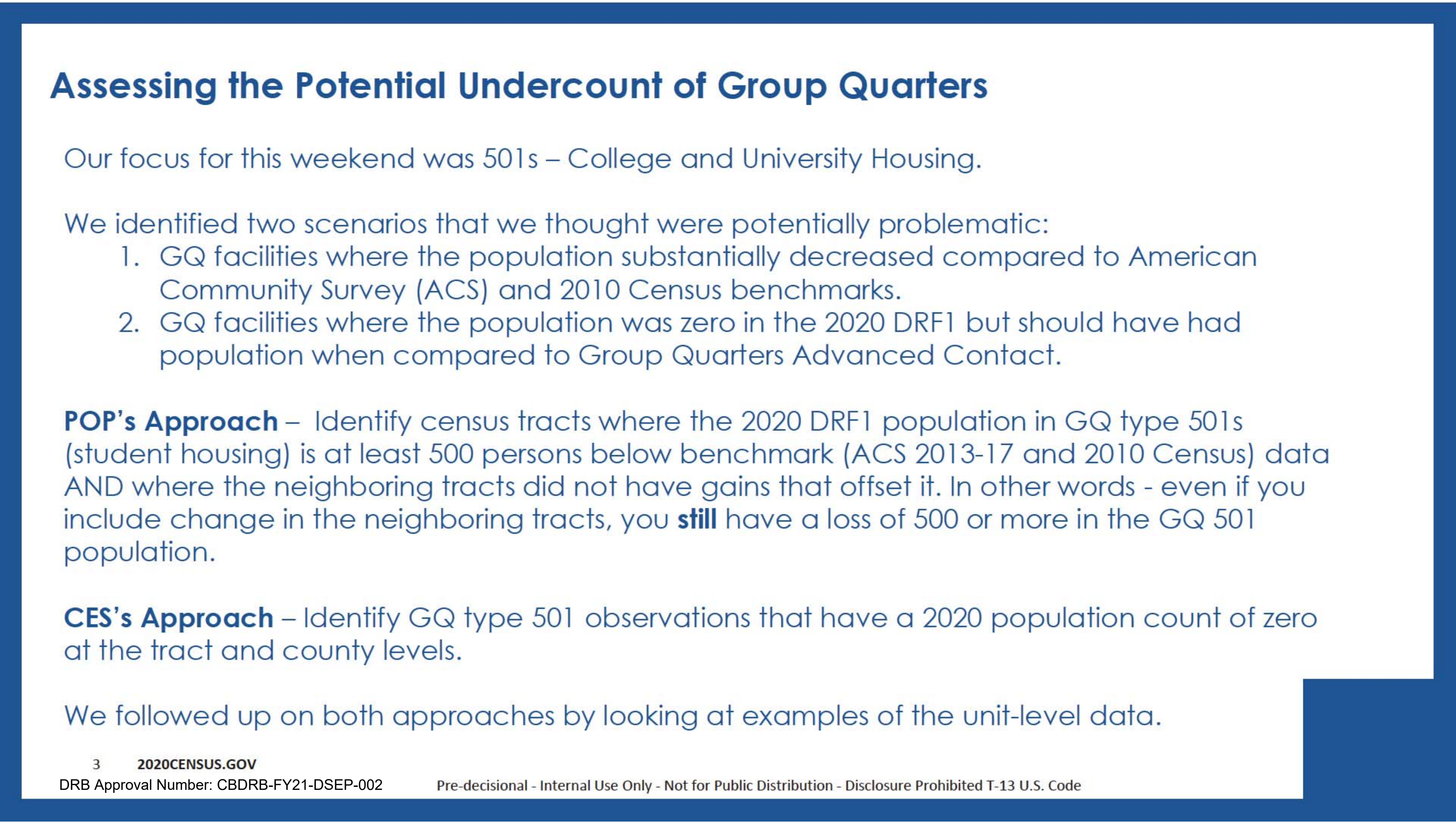
DRB Approval Number: CBDRB-FY21-DSEP-002



Group Quarters (GQ) Review – the Undercount Issue

The request for us this weekend was to assess the magnitude of the potential undercount of GQ persons and whether this warrants additional action.

The primary challenge faced while tackling this issue is that the **units within a GQ facility are not linked**. Because of this, there is no one systematic criteria that can be implemented in order to identify every potentially problematic GQ. POP and CES worked together to come up with two independent approaches that get at the universe for this issue. GEO and DSSD also conducted their own analysis.



Assessing the Potential Undercount of Group Quarters

Our focus for this weekend was 501s – College and University Housing.

We identified two scenarios that we thought were potentially problematic:

1. GQ facilities where the population substantially decreased compared to American Community Survey (ACS) and 2010 Census benchmarks.
2. GQ facilities where the population was zero in the 2020 DRF1 but should have had population when compared to Group Quarters Advanced Contact.

POP's Approach – Identify census tracts where the 2020 DRF1 population in GQ type 501s (student housing) is at least 500 persons below benchmark (ACS 2013-17 and 2010 Census) data AND where the neighboring tracts did not have gains that offset it. In other words - even if you include change in the neighboring tracts, you **still** have a loss of 500 or more in the GQ 501 population.

CES's Approach – Identify GQ type 501 observations that have a 2020 population count of zero at the tract and county levels.

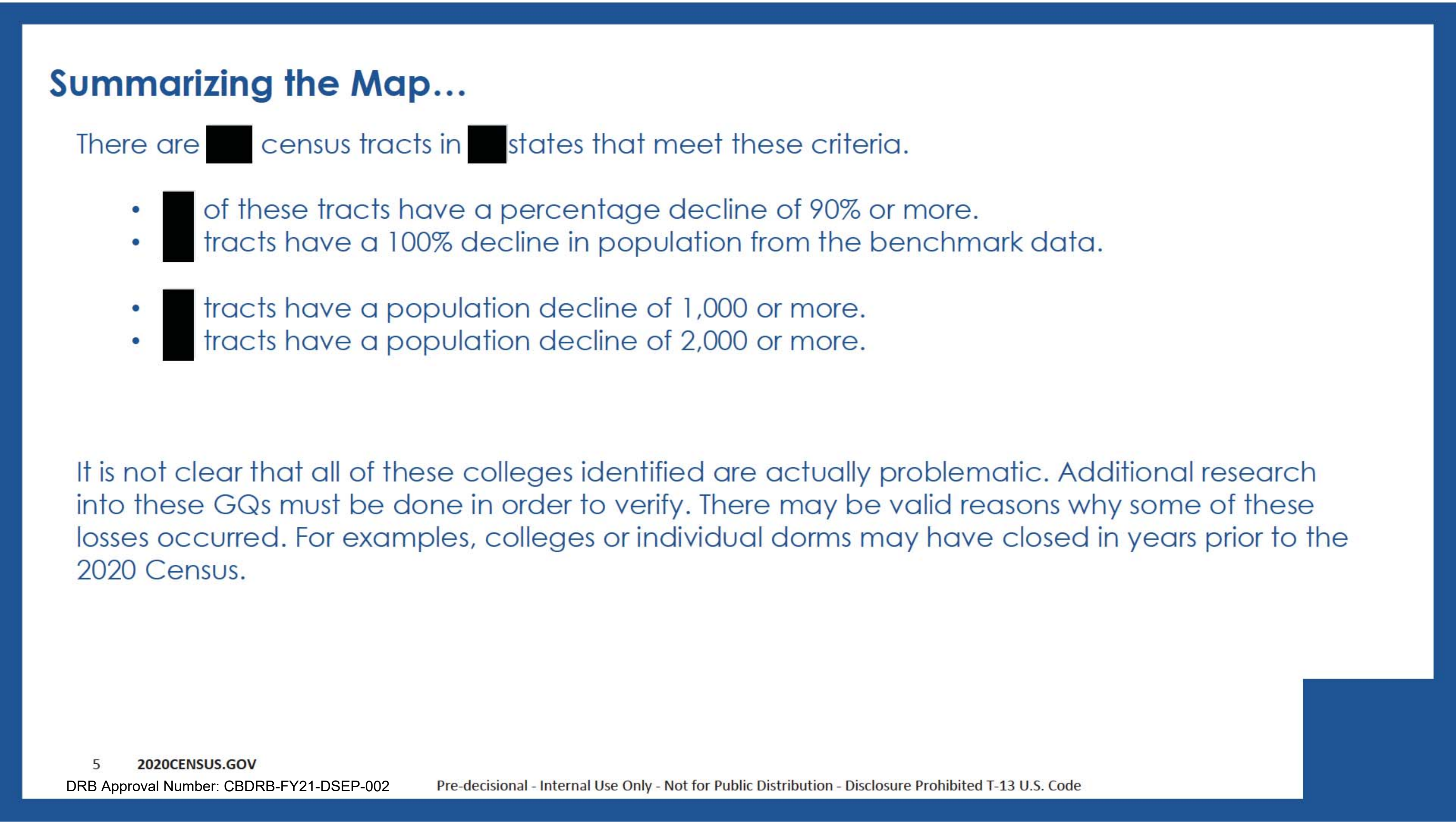
We followed up on both approaches by looking at examples of the unit-level data.

DRB Approval Number: CBDRB-FY21-DSEP-002

Pre-decisional - Internal Use Only - Not for Public Distribution - Disclosure Prohibited T-13 U.S. Code

POP’s Approach for GQ Type 501 (College/University Housing)

[redacted]



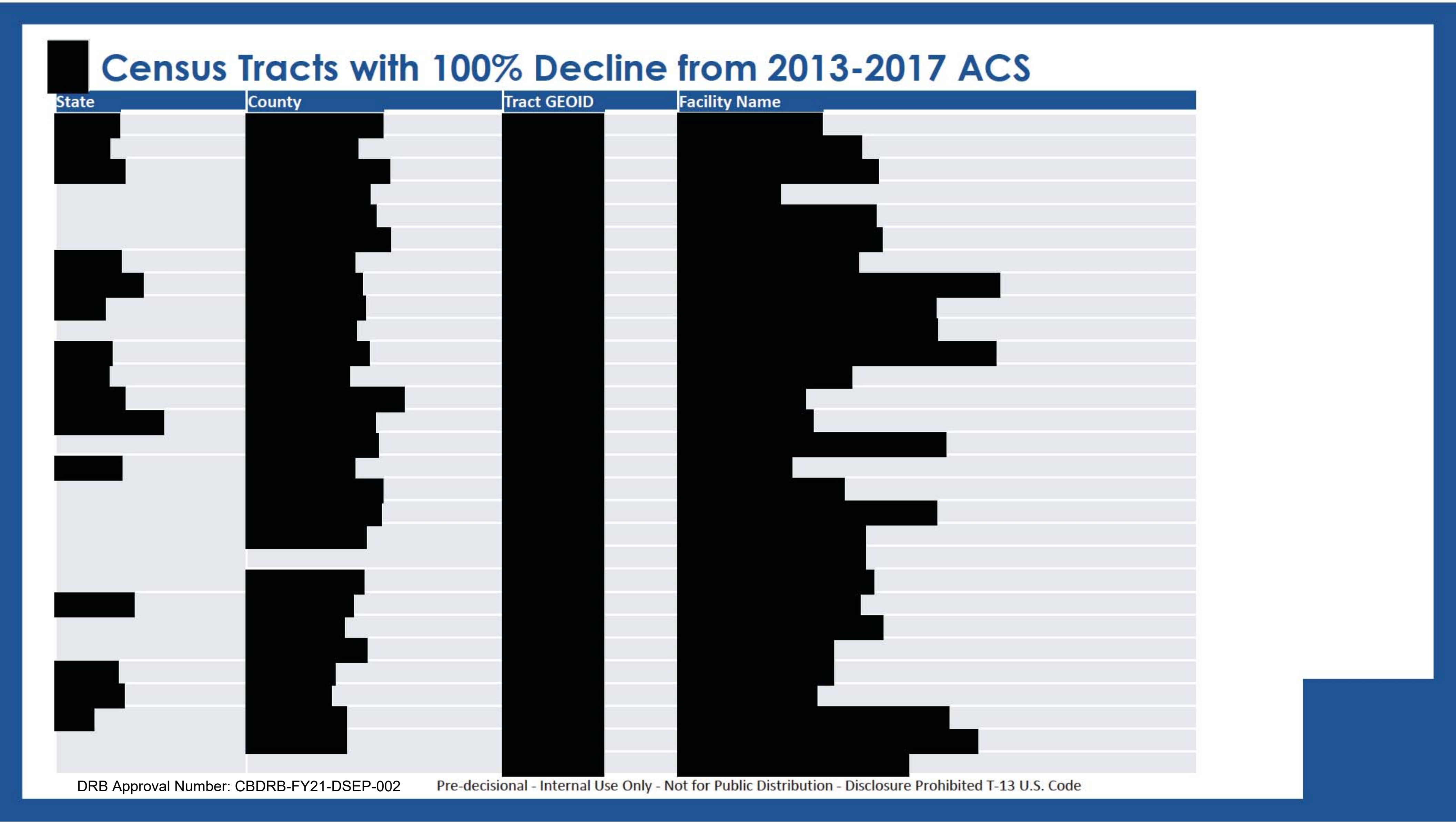
Summarizing the Map…

There are [redacted] census tracts in [redacted] states that meet these criteria.

- [redacted] of these tracts have a percentage decline of 90% or more.
- tracts have a 100% decline in population from the benchmark data.

- [redacted] tracts have a population decline of 1,000 or more.
- tracts have a population decline of 2,000 or more.

It is not clear that all of these colleges identified are actually problematic. Additional research into these GQs must be done in order to verify. There may be valid reasons why some of these losses occurred. For examples, colleges or individual dorms may have closed in years prior to the 2020 Census.



[redacted]

Census Tracts with 100% Decline from 2013-2017 ACS

State	County	Tract GEOID	Facility Name
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
		[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
		[redacted]	[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002

Pre-decisional - Internal Use Only - Not for Public Distribution - Disclosure Prohibited T-13 U.S. Code

College with [redacted] Population Loss

[redacted]

[redacted]

[redacted]

[redacted]

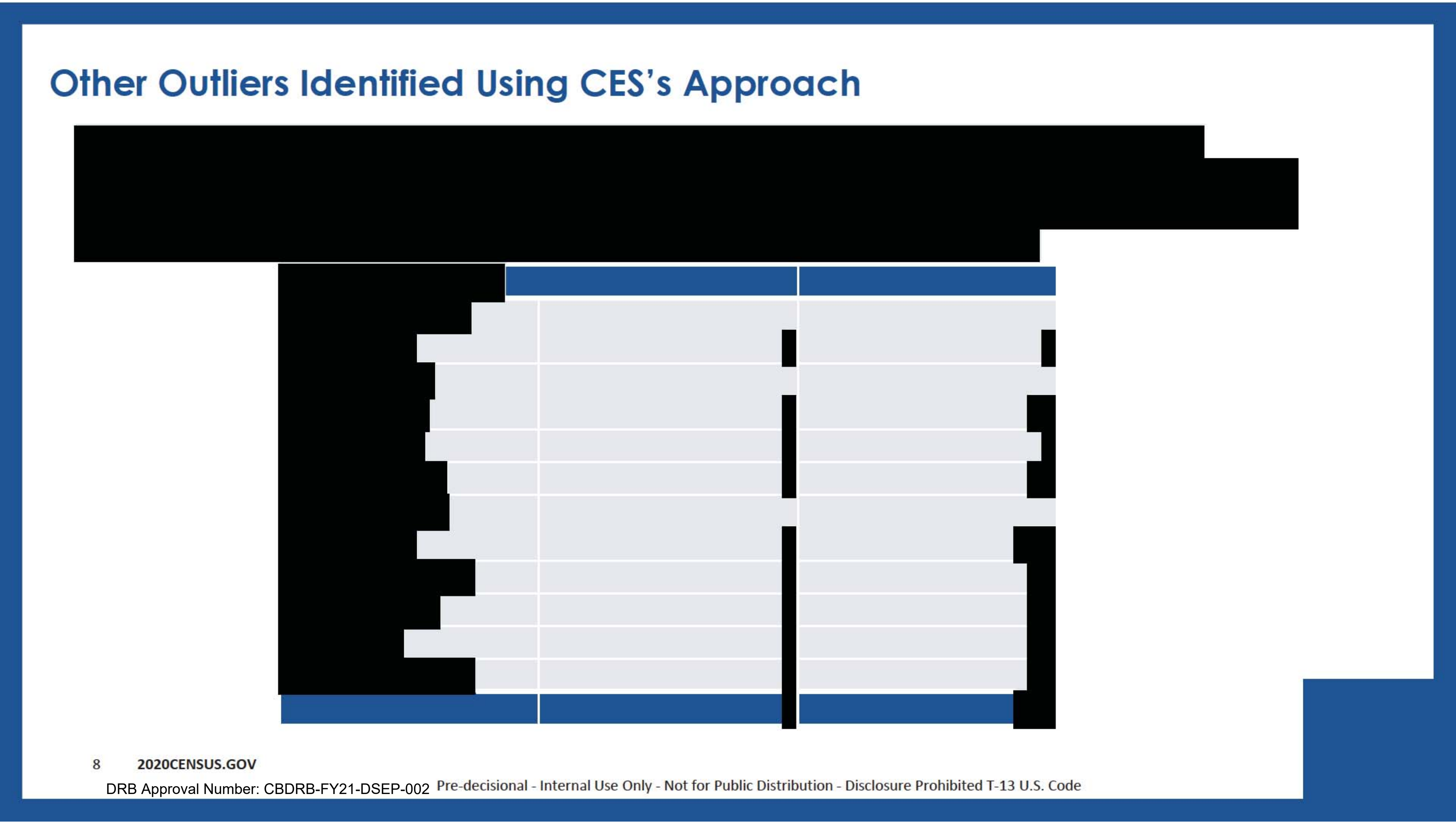
[redacted]

GQ Unit Name	2020 DRF1 POP	Expected Pop
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002

[redacted]

de



Other Outliers Identified Using CES's Approach

[redacted]

[redacted]		
[redacted]		
[redacted]	[redacted]	[redacted]
[redacted]		
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]		
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002 Pre-decisional - Internal Use Only - Not for Public Distribution - Disclosure Prohibited T-13 U.S. Code

POP's Approach – Prisons – [redacted] Census Tracts

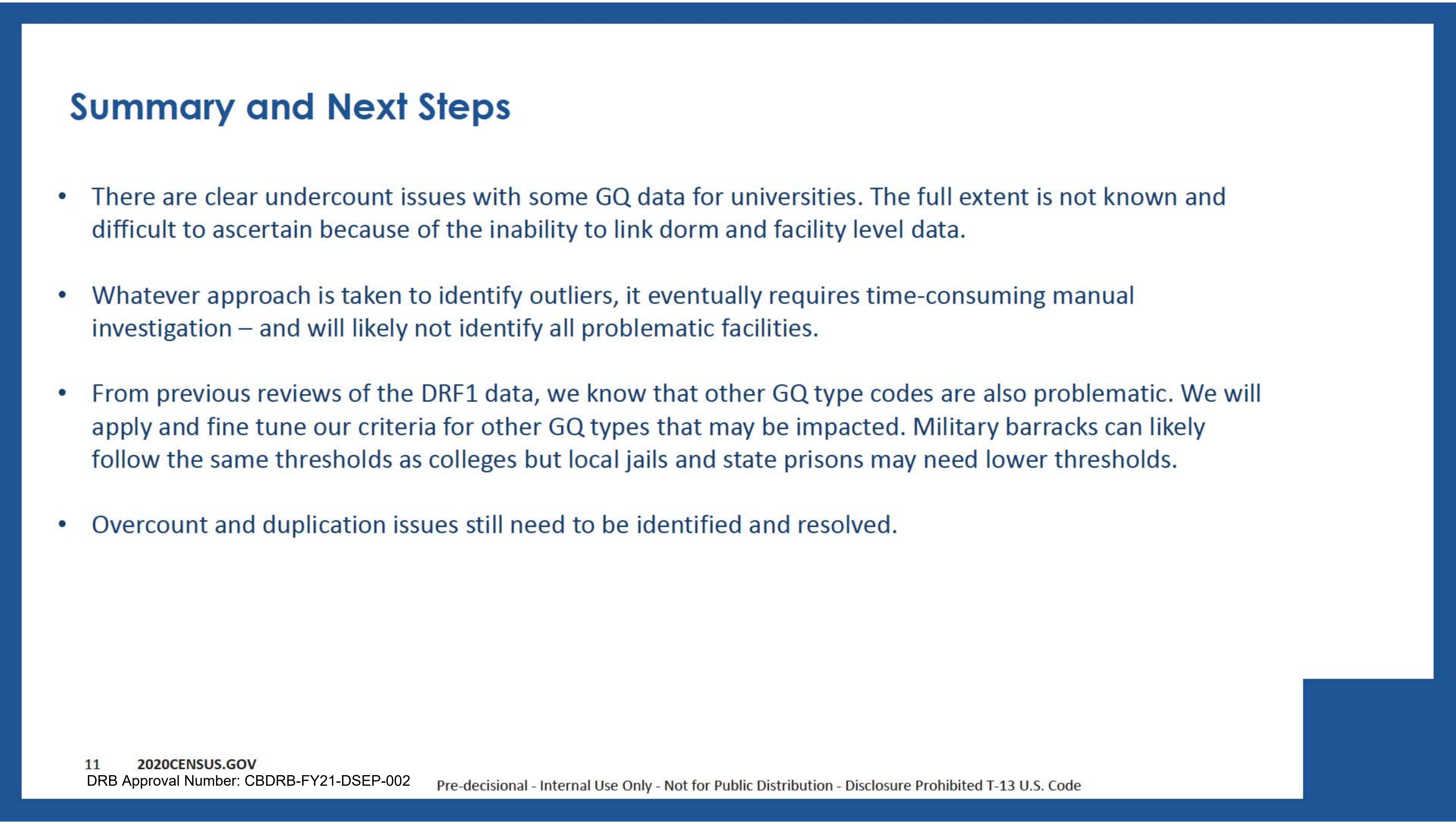
Group Quarters Anomalies: Type 100s (Adult Correctional Facilities)

[redacted]

POP's Approach – Military Quarters – [redacted] Census Tracts

Group Quarters Anomalies: Type 600s (Military Quarters)

[redacted]



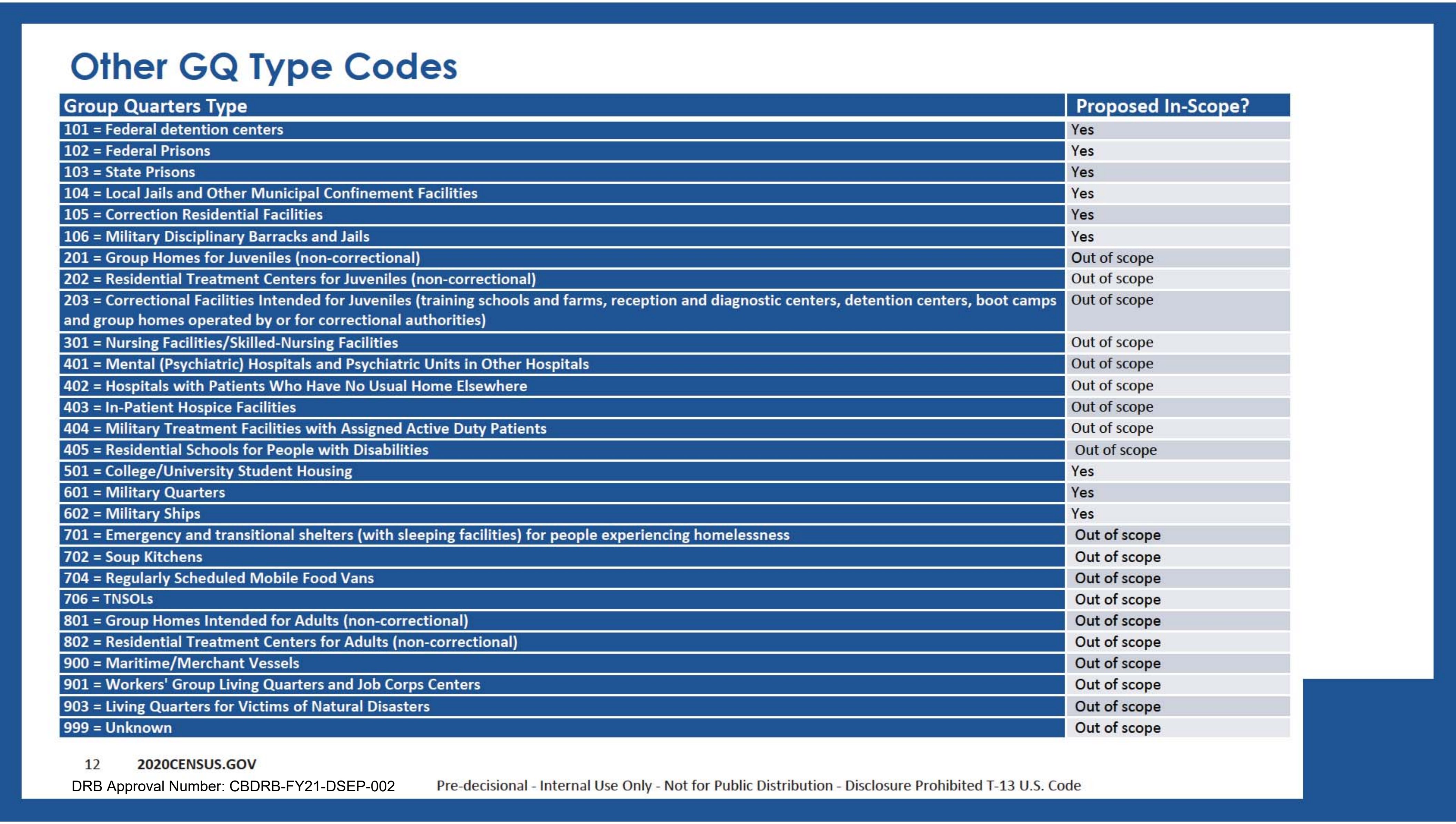
Summary and Next Steps

- There are clear undercount issues with some GQ data for universities. The full extent is not known and difficult to ascertain because of the inability to link dorm and facility level data.

- Whatever approach is taken to identify outliers, it eventually requires time-consuming manual investigation – and will likely not identify all problematic facilities.

- From previous reviews of the DRF1 data, we know that other GQ type codes are also problematic. We will apply and fine tune our criteria for other GQ types that may be impacted. Military barracks can likely follow the same thresholds as colleges but local jails and state prisons may need lower thresholds.

- Overcount and duplication issues still need to be identified and resolved.



Other GQ Type Codes

Group Quarters Type	Proposed In-Scope?
101 = Federal detention centers	Yes
102 = Federal Prisons	Yes
103 = State Prisons	Yes
104 = Local Jails and Other Municipal Confinement Facilities	Yes
105 = Correction Residential Facilities	Yes
106 = Military Disciplinary Barracks and Jails	Yes
201 = Group Homes for Juveniles (non-correctional)	Out of scope
202 = Residential Treatment Centers for Juveniles (non-correctional)	Out of scope
203 = Correctional Facilities Intended for Juveniles (training schools and farms, reception and diagnostic centers, detention centers, boot camps and group homes operated by or for correctional authorities)	Out of scope
301 = Nursing Facilities/Skilled-Nursing Facilities	Out of scope
401 = Mental (Psychiatric) Hospitals and Psychiatric Units in Other Hospitals	Out of scope
402 = Hospitals with Patients Who Have No Usual Home Elsewhere	Out of scope
403 = In-Patient Hospice Facilities	Out of scope
404 = Military Treatment Facilities with Assigned Active Duty Patients	Out of scope
405 = Residential Schools for People with Disabilities	Out of scope
501 = College/University Student Housing	Yes
601 = Military Quarters	Yes
602 = Military Ships	Yes
701 = Emergency and transitional shelters (with sleeping facilities) for people experiencing homelessness	Out of scope
702 = Soup Kitchens	Out of scope
704 = Regularly Scheduled Mobile Food Vans	Out of scope
706 = TNSOLs	Out of scope
801 = Group Homes Intended for Adults (non-correctional)	Out of scope
802 = Residential Treatment Centers for Adults (non-correctional)	Out of scope
900 = Maritime/Merchant Vessels	Out of scope
901 = Workers' Group Living Quarters and Job Corps Centers	Out of scope
903 = Living Quarters for Victims of Natural Disasters	Out of scope
999 = Unknown	Out of scope

DRB Approval Number: CBDRB-FY21-DSEP-002

Pre-decisional - Internal Use Only - Not for Public Distribution - Disclosure Prohibited T-13 U.S. Code

Thanks!

POP’s Approach for GQ Type 501 90%+

Group Quarters Anomalies: Type 501 (College/University Housing)

[redacted]

POP's Approach for GQ Type 501 – 100%+

Group Quarters Anomalies: Type 501 (College/University Housing)

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002

Pre-decisional - Internal Use Only - Not for Public Distribution - Disclosure Prohibited T-13 U.S. Code

EGG Presentation Recap

- It went very well. Senior management agreed with the importance of identifying and fixing these GQ issues. Marc is taking the lead on organizing.
- Concurrent paths moving forward to identify the general magnitude of problematic GQ counts:
  - **GEO** is going to add their spatial expertise, looking at tract-level data to see where 2020 GQ populations don't align with the benchmark data. The initial focus for GEO will be on identifying situations where the 2020 GQ population in DRF1 is lower than expected, since we don't expect this situation to improve in subsequent DRF2 processing.
  - **DSSD** will begin doing a name matching exercise for the GQ universe. This will help identify GQs where the DRF1 count is higher than benchmark where duplication is suspected.
  - **POP** will continue doing a first pass at tract-level maps of GQ population data compared with benchmarks, tabulating a list of GQ facilities by state with populations that differ substantially.
- **Potential role of FSCPE:** They could review the draft list of potential GQ problems and provide expertise as planned for the GQ count review operation.
- Knowing the rough magnitude of the problem will then help us identify feasible solutions.
- The Data Quality EGG will meet again at 930 am this morning to continue the discussion.

Andrew Keller
Imputing GQ Pop Counts – Draft 1
December 6, 2020

Table 1: Input Data

	No Good Person	Has Good Person	Total
Occupied GQ	18,646	180,396	199,042
Delete GQ	7,225	381	7,606
Nonresidential GQ	2,373	76	2,449
Vacant During Visit, Open on Census Day	19,683	1,542	21,225
Refusal GQ	6,756	973	7,729
Vacant GQ	29,229	968	30,197
Total	83,912	184,336	268,248

1. Red and Green (223,163 cases)
    a. These are the resolved cases – use appropriate count
    b. Red are the donors on the models below
2. Blue (45,085 cases) – These are the unresolved cases. We believe them to be occupied, but do not have a good person count.

Business Rules

1. If the unresolved cases was a GQ in 2010 and had a pop count, I am going to directly assign that pop count.
2. If not, I use the modeled result.

Two Models

1. Has 2020 GQ Expected Count - Linear Regression Model
    a. DV: ratio of 2020 Good Person Count / 2020 GQ Expected Count
    b. 91,658 of the 180,396 cases have 2020 GQ Expected Count
    c. Score model over 9,020 unresolved cases with a 2020 GQ Expected Count. This outputs an estimated occupied ratio which I multiply by the 2020 GQ Expected count to get an imputed GQ count.
2. No 2020 GQ Expected Count - Linear Regression Model
    a. DV: 2020 Good Person Count
    b. 88,738 of the 180,396 cases without 2020 GQ Expected Count
    c. Score model over 36,065 unresolved cases without a 2020 GQ Expected Count. This outputs an imputed GQ count.

Results

1. Use 2020 ACS GQ Count As a Baseline
2. Compare Results Between No Imputation (Keeping a 0 for all Blue Cases) and Imputation (Applying Business Rules and Models

2020 ACS GQ Count – 8,084,362
2020 Census GQ Count (No Imputation) – 8,294,160
2020 Census GQ Count (With Imputation) – 10,198,552

Path	GQ	% of GQ	GQ People	% of GQ People
Resolved	223,163	83.2%	8,294,160	81.3%
Has 2020 Expected Pop, Use 2010 GQ Count	5,650	2.1%	252,257	2.5%
Has 2020 Expected Pop, Use Model	3,370	1.3%	300,883	3.0%
Without 2020 Expected Pop, Use 2010 GQ Count	11,393	4.2%	462,162	4.5%
Without 2020 Expected Pop, Use Model	24,672	9.2%	889,090	8.7%
Total	268,248	100.0%	10,198,552	100.0%

12/6/20 – Models being refined

2020 GQ Count (No Imputation Ratio) – 1.03
2020 GQ Count (With Imputation) – 1.26

Model Appendix

1. Has 2020 GQ Expected Count

```
proc reg data=yesmaxmod outest=yesmaxparam;
     model filledratio = /* feddc */ statejail localjail housejail nursing
college military homeless soup /* uaa */ group dne2010 ar1 ar2 ar3 ar6 max5l
max1\
00m nomax;
run;
```

2. No 2020 GQ Expected Count

```
proc reg data=nomaxmod outest=nomaxparam;
     model gp = feddc statejail  localjail housejail nursing college military
homeless soup uaa group dne2010 ar1 ar2 ar3 ar6 max5l max100m nomax;
run;
```

County Distribution of 2020 Census / 2020 ACS - GQ Person Ratios Before Imputation

[redacted]

County Distribution of 2020 Census / 2010 ACS - GQ Person Ratios After Imputation

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

[redacted]

[redacted]

[redacted]

[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

[redacted]

[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

[redacted]

[redacted]

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002

[redacted]

[redacted]

[redacted]

PreDecisional Information for Internal Use OnlyDraft results still being reviewed. Title 13 Data Results have not been through disclosure avoidance

Record Linkage Analysis of Group Quarters

Tom Mule 12/6/2020 Draft

Question: Can computer matching be used to find duplicate group quarters people in different MAFIDs?

2020 DRF1 GQ Person records

Conducted record linkage using the 2010 Census DPI code.

- matching code that searched for duplicates of census people to other MAFIDs in 2010 Census
- used for 2020 SRQA matching
- research leading to 2010 developed matching cutoffs based on last name frequencies and geographic distance apart
- Geographic distance: Developed for 2010 Census Collection blocks that have 5 characters. The 2020 BCUs have 8 characters. This analysis used spaces 3 to 7 of BCU 00234500 to be the same "BCU subset"
- Matched all of the DRF1 GQ Person
- Showing matching results for Good Persons who are data-defined. My results may be different than others and are not official tabulations.

How many good links were made to different GQ matching IDs?

Table 1: Person Links found between different GQ Matching IDs

Links			
Within BCU subset (3 to 7)	Outside BCU subset, Within Tract	Outside Tract, Within County	Different County, Same State
211,000	33,00	28,000	17,000

Note: BCU subset is values in spaces 3 to 7

Links are "edges"
If person 1 and 2 are duplicates there are one link: 1 to 2
If person 1, 2 and 3 are duplicates then there are 3 links: 1 to 2, 1 to 3 and 2 to 3.
If person 1, 2, 3 and 4 are duplicates then there are 6 links: 1 to 2, 1to3, 1to4, 2to3,2to4,3to4
Etc.

PreDecisional Information for Internal Use OnlyDraft results still being reviewed. Title 13 Data
Results have not been through disclosure avoidance

What is the match rate for the GQ Mafids? How many of the people in GQ mafid are matching?

For this analysis, if a MAFID had multiple SolicitationID for the response, I picked the solicitation ID with the most number of good data-defined people.

- This subset was done for this quick analysis and is something I did for this quick analysis
- This analysis has 183,000 MAFIDs and 8,649,000 person records
- Numerator is number of persons who have a match
- Denominator number of good data-defined person records in the selected solicitation ID for MAFID

Table 2: Distribution of Match Rates for Selected Records in GQ MAFIDs

Match Rate Interval	Number of MAFIDs	Person Records
0	169,000	6,784,000
>0 to <.5	10,000	1,682,000
.5 to .75	900	41,000
>.75 to .99	700	96,000
>.99 to 1	1,800	45,000

If group quarters A has 100 people and group quarters B has 100 and there are all the same people then this would show up as 2 MAFIDs and 200 person records in the >.99 to 1 row

Group Quarters Overcoverage Research Update

Tom Mule DSSD December 23, 2020 Draft

DSSD has been researching duplication of Persons in GQ mafids to other GQ mafids to assess magnitude. Research has led to the following remedy for implementation consideration.

Used Test DRF2 Data

- This test data already accounts for primary selection and other preplanned processing steps
- Our concern is that the preplanned steps will not account for people being enumerated at multiple GQ MAFIDs

Conducted record linkage using the 2010 Census DPI code.

- matching code that searched for duplicates of census people to other MAFIDs in 2010 Census
- used for 2020 SRQA matching
- research leading to 2010 developed matching cutoffs based on last name frequencies and geographic distance apart. (Ikeda and Porter (2008))
- Work used only duplicates found within the same state
- Geographic distance: Decisions account for whether links are within same BCU, same tract, same county or different county

2010 DPI code could identify Matches and Possible Matches based on the 2010 research criteria

- Used the Matches within State
- Possible Matches
  - 2010 Census Coverage Measurement had possible matches since there was a clerical matching step to review those possible matches
  - We do not have a clerical matching step
  - Reviewed possible match to apply additional rules for usage
  - Link had to be within tract
  - First and Last Name had to have exact agreement
  - Month and Day of birth had to match exactly or be missing on both
- Concern about making matches based on placeholder names
  - matches with response names that began with FIRST, LAST, RESIDENT, PERSON, STUDENT, WARD, BED, COED, UNABLE or DSS were not used
  - Only use matches within the same group quarters type (first digit of GQ type code) or one of the persons was in a Workers' Group Living Quarters and Job Corps Centers (901)

Computer matching results

- 424,000 GQ persons identified duplicated to another GQ mafid
  - Person is in multiple GQ mafids
  - Person is in only one of 202,000 groupings
  - Processing can try to identify which record to keep
  - 202,000 good persons who stay in census
  - 222,000 possible duplicates who should not be counted

Number of person records in each person duplication group	Percent
2	93%
3	6%
4 or more	1%

For each of the 202,000 groupings, algorithm to pick which record to remove and which to keep.

Records at the end of the sort are ones to keep

1. How does the GQ count compare to the Maximum reported in advanced contact?
   a. Higher
   b. Missing
   c. Lower than or equal
2. How high is the GQ count?
   a. Records would be removed from GQs with higher counts
3. How does the GQ count compare to the Expected reported in advanced contact?
   a. Higher
   b. Missing
   c. Lower than or equal

Here is an example

Example 1: Rule Application #1

MAFID	Person	Maximum reported	GQ record count	Expected reported	Decision
1	1	25	50	20	Remove
2	2	Not reported	20	Not reported	Remove
3	3	100	75	60	Remove
4	4	100	75	80	Keep

- Person 1 is removed because this group quarters has more people than the maximum count provided during advanced contact

- Person 2 is removed because this group quarters did not participate in the advance contact. We have another group quarters that did so will continue on

- Person 3 and Person 4 have the same gq count so they are tied on the second sort. Person 3 is removed because MAFID 3 has a count higher than expected while MAFID 4 is under their expected count.

Additional rule for unduplication for these 202,000 person groups

- 6,000 MAFIDs had only one duplicate person removed based on the above algorithm
- Concern that these could be false matches
- To mitigate, we modified that we would only implement this for GQ MAFIDs that had two or more duplicates removed.

424,000 Person records in 202,000 Person Groups

- 222,000 persons identified for initial removal
- 6,000 were in MAFIDs where only person being removed
- 216,000 person records were identified for final removal

Table 1: Summation by Groupings of Number of Duplicate Records Removed in each GQ MAFID

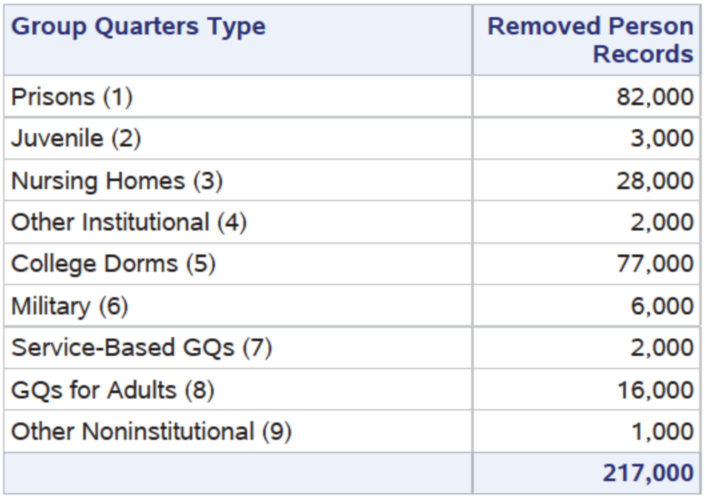
Duplicate Persons Removed From the GQ MAFID	Number of GQ MAFIDs	Duplicate Persons Removed
2 People	2,000	4,000
3 to 5 People	2,000	8,000
6 to 10 People	1,000	8,000
11-100 People	2,000	61,000
101-999 People	450	109,000
1,000 or more	<15	26,000
Total	7,000	216,000

Note: Numbers may not add and may be different due to rounding individual rows

Table 2 shows a research initial implementation of the rules to show the magnitudes for group quarters types.

Table 2: Unduplication Results by Group Quarters Facilities Type

Note: Numbers may not add or may be different due to rounding different individual rows.



Group Quarters Type	Removed Person Records
Prisons (1)	82,000
Juvenile (2)	3,000
Nursing Homes (3)	28,000
Other Institutional (4)	2,000
College Dorms (5)	77,000
Military (6)	6,000
Service-Based GQs (7)	2,000
GQs for Adults (8)	16,000
Other Noninstitutional (9)	1,000
	217,000

Continuing Analysis

DSSD has reached out to Population Division to start sharing these results.

DSSD has been asked to examine if there is any overlap between GQ Unduplication MAFIDs and Calling Results File. The unduplication procedure described here may be modified based on those findings.

Implementation

DSSD would conduct the person matching and processing to identify the person records to be removed.

DSSD would deliver the person records to be removed to DRPS

DSSD has sent a Memorandum of Understanding to DRPS about implementing this removal at the same time that DRPS is doing the processing to add the additional people identified by the undercoverage operations.

PreDecisional information for internal usage only   Draft results still being reviewed   Title 13 counts Do Not disclose No Disclosure Avoidance has been applied
Not Official Results and may differ from other results

	Group Quarters Types (first digit of GQtype of MAFID)																		All	
	Jails (100s)		Juvenille (200s)		Nursing Homes(300s)		Treatment (400s)		College Dorms (500s)		Military (600s)		Treatment (700s)		GQs for Adults, Residential(800s)		Others(900s)			
	scount		scount		scount		scount		scount		scount		scount		scount		scount		scount	
	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum
match_rate																				
0	9,600	1,050,000	5,500	72,000	20,000	1,157,000	1,700	44,500	27,000	2,104,000	2,500	209,000	23,000	436,000	58,000	584,000	15,500	183,000	162,000	5,840,000
>0 to <.5	3,300	948,000	500	16,500	4,300	431,000	200	20,500	3,000	667,000	500	147,000	1,300	99,000	2,600	95,500	400	27,000	16,000	2,451,000
.5 to .75	100	17,500	70	950	90	3,900	<15	300	200	22 500	20	650	100	1,600	600	6,800	80	1,800	1,200	56,000
>0.75 to .99	150	37,000	50	1,600	200	14,500	20	1,800	200	35,000	<15	900	<15	400	450	8,900	30	750	1 100	101,000
>.99 to 1	90	15,000	150	1,500	300	11,500	30	950	350	17,500	40	300	<15	<15	1,000	6,800	150	1,300	2,100	55,000
All	13,000	2,068,000	6,200	92,500	24,500	1,619,000	1,900	67,500	30,500	2,846,000	3,100	358,000	24,500	537,000	62,500	702,000	16,000	214,000	183,000	8,504,000

My initial processing has 183,000 GQs MAFIDS with 8,504,000 GQ "good person" records

Table shows for each GQ facility the rate of good person records who were computer linked to another person to another group quarters MAFID

- Matches and also including possible matches within the tract. Possible matches are cases that agree on atleast first name, last name and sex but do not meet match criteria for tract.

99 to 1 is everyone or close to all people are found in another group quarters   N=number of MAFIDs. Sum=sum of good persons in those MAFIDs

- 2,100 MAFIDs with 55,000 people
- More analysis can be done today about how to unduplicate them
- College Dorms is now down to 350 MAFIDs after applying the patch yesterday.
- The table shows results for the other types

>.75 to .99 is the next band where over ¾ of the people are found in another group quarters

- 1,100 MAFIDs with 101,000 people
- Potential another set of Group Quarters MAFIDs where some of the population could be unduplicated

Continuing to do analysis to check the matching done so far

PreDecisional information for internal usage only Draft results still being reviewed Title 13 counts Do Not disclose No Disclosure Avoidance has been applied
Not Official Results and may differ from other results

	Group Quarters Types (first digit of GQtype of MAFID)																		All	
	Jails (100s)		Juvenille (200s)		Nursing Homes(300s)		Treatment (400s)		College Dorms (500s)		Military (600s)		Treatment (700s)		GQs for Adults, Residential(800s)		Others(900s)			
	scount		scount		scount		scount		scount		scount		scount		scount		scount		scount	
	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum
match_rate																				
0	9,600	1,050,000	5,500	72,085	20,000	1,157,000	1,700	44,500	27,000	2,104,000	2,500	209,000	23,000	436,000	58,000	584,000	15,500	183,000	162,000	5,840,000
>0 to <.5	3,300	948,000	500	16,500	4,300	431,000	200	20,500	3,000	667,000	500	147,000	1,300	99,000	2,600	95,500	400	27,000	16,000	2,451,000
.5 to .75	100	17,500	70	950	90	3,900	<15	300	200	22,500	20	600	100	1,600	600	6,800	80	1,800	1,200	56,000
>0.75 to .99	150	37,000	50	1,600	200	14,500	20	1,800	200	35,000	<15	900	<15	400	450	8,900	30	750	1,100	101,000
>.99 to 1	90	15,000	150	1,500	300	12,000	30	950	300	17,500	40	300	<15	<15	1000	6,800	150	1,300	2,100	55,000
All	13,000	2,068,000	6,200	92,500	24,500	1,619,000	2,000	67,500	30,500	2,846,000	3,100	358,000	24,500	538,000	62,500	702,000	16,000	214,000	183,000	8,504,000

My initial processing has 183,000 GQs MAFIDS with 8,504,000 GQ "good person" records

Table shows for each GQ facility the rate of good person records who were computer linked to another person to another group quarters MAFID

- Matches and also including possible matches within the tract. Possible matches are cases that agree on atleast first name, last name and sex but do not meet match criteria for tract.

99 to 1 is everyone or close to all people are found in another group quarters N=number of MAFIDs. Sum=sum of good persons in those MAFIDs

- 2,100 MAFIDs with 55,000 people
- More analysis can be done today about how to unduplicate them
- College Dorms is now down to 350 MAFIDs after applying the patch yesterday.
- The table shows results for the other types

>.75 to .99 is the next band where over ¾ of the people are found in another group quarters

- 1,100 MAFIDs with 101,000 people
- Potential another set of Group Quarters MAFIDs where some of the population could be unduplicated

Continuing to do analysis to check the matching done so far. The rest of this document looks further at the 99 or higher row to examine tracts where there is large happenings.

Statistics have been rounded according to Census Bureau disclosure standards

These are the tracts with the GQ Prison 100s types.
It looks at the 25 largest BCU geography tracts where 99 percent of the MAFID matches to another GQ.

One thing that stands out is that several of these tracts have only one GQ MAFID.

- In looking at [redacted], this MAFID has almost everyone in another GQ in a different tract.

Table x: BCU Geography Tracts with Largest number of Persons in MAFIDs with 99+ percent match rate

Obs	BCUSTATEFP	BCUCOUNTYFP	BCUTRACTCE	_FREQ_	sum_scount
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	3,100
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	950
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	900
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	850
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	750
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	650
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	350
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	350
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	350
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	**14,000**

This is the 25 largest tracts when doing for Nursing Home 300s

These are MAFIDs where over 99 percent are found in another group quarters

Similar results as prisons. These tracts have only one MAFID. When this is happening that the people are found in a MAFID that is in a different tract.

Table 4: 25 Tracts with Largest Number of Nursing Home People Found in a Group Quarters

Obs	BCUSTATEFP	BCUCOUNTYFP	BCUTRACTCE	_TYPE_	_FREQ_	sum_scount
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	100
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	100
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	100
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	4,800

This is College Dorms GQ Types 500s

Table 4: Largest 25 Tracts for College Dorms GQ 500s where over 99 percent of the people in the MAFD match to another GQ

These could possibly be candidates for doing what was just done for the previous patch.

Obs	BCUSTATEFP	BCUCOUNTYFP	BCUTRACTCE	_TYPE_	_FREQ_	sum_scount
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	350
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	13,000

PreDecisional information for internal usage only Draft results still being reviewed Title 13 counts Do Not disclose No Disclosure Avoidance has been applied
Not Official Results and may differ from other results 12/09/20 Draft

	Group Quarters Types (first digit of GQtype of MAFID)																		All	
	Jails (100s)		Juvenille (200s)		Nursing Homes(300s)		Treatment (400s)		College Dorms (500s)		Military (600s)		Treatment (700s)		GQs for Adults, Residential(800s)		Others(900s)			
	scount		scount		Scount		scount		scount		scount		scount		scount		scount		scount	
	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum	N	Sum
match_rate																				
0	7,000	650,000	5,000	62,000	16,500	917,000	1,400	31,000	24,500	1,721,000	2,200	165,000	21,500	361,000	53,500	487,000	15,000	165,000	147,000	4,559,000
>0 to <.5	5,100	1,268,000	750	22,500	6,800	642,000	400	31,000	4,500	968,000	750	188,000	2,800	171,000	5,000	170,000	700	41,000	27,000	3,502,000
.5 to .75	300	25,500	100	1,600	150	8,100	30	500	350	48,000	20	800	250	4,100	1,100	11,500	150	3,500	2,500	103,000
>0.75 to .99	500	86,000	100	3,300	400	30,000	40	3,000	400	65,500	30	2,400	30	1,400	900	18,000	70	1,700	2,500	211,000
>.99 to 1	350	38,500	250	2,800	550	22,000	60	2,000	700	44,000	60	1,400	<15	20	2,000	14,500	350	3,000	4,400	128,000
All	13,000	2,068,000	6,200	92,500	24,500	1,619,000	1,900	67,500	30,500	2,846,000	3,100	358,000	24,500	537,000	62,500	702,000	16,000	214,000	183,000	8,504,000

My initial processing has 183,000 GQs MAFIDS with 8,504,000 GQ "good person" records. This 12/09/20 version corrects for an error from the 12/08/20 verison.

Table shows for each GQ facility the rate of good person records who were computer linked to another person to another group quarters MAFID

- Matches and also including possible matches within the tract. Possible matches are cases that agree on atleast first name, last name and sex but do not meet match criteria for tract.

99 to 1 is everyone or close to all people are found in another group quarters N=number of MAFIDs. Sum=sum of good persons in those MAFIDs

- 4,400 MAFIDs with 128,000 people
- If MAFID A has 100 people
- More analysis can be done today about how to unduplicate them.
- College Dorms is now 700 MAFIDs after applying the patch yesterday.
- The table shows results for the other types

>.75 to .99 is the next band where over ¾ of the people are found in another group quarters

- 2,500 MAFIDs with 211,000 people
- Potential another set of Group Quarters MAFIDs where some of the population could be unduplicated

Continuing to do analysis to check the matching done so far. The rest of this document looks further at the 99 or higher row to examine tracts where there is large happenings.

These are the tracts with the GQ Prison 100s types.
It looks at the 25 largest BCU geography tracts where 99 percent of the MAFID matches to another GQ.

One thing that stands out is that several of these tracts have only one GQ MAFID.
The 12/09/20 correction is showing that the top 25 now generally have 2+ MAFIDs in the tract.

Sum_count is the number of people in the MAFIDs in the tact that have 99+ percent match rate

Table x: BCU Geography Tracts with Largest number of Persons in MAFIDs with 99+ percent match rate

Obs	BCUSTATEFP	BCUCOUNTYFP	BCUTRACTCE	_TYPE_	_FREQ_	sum_scount
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	5,600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	2,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,900
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,800
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	950
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	900
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	850
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	750
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	750
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	650
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	650
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	550
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	**29,000**

This is the 25 largest tracts when doing for Nursing Home 300s

These are MAFIDs where over 99 percent are found in another group quarters

Similar results as prisons. The 12/09/20 rerun is finding a second MAFID in the tract.

Table 4: 25 Tracts with Largest Number of Nursing Home People Found in a Group Quarters

Obs	BCUSTATEFP	BCUCOUNTYFP	BCUTRACTCE	_TYPE_	_FREQ_	sum_scount
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	650
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	550
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	550
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	8,300

This is College Dorms GQ Types 500s

Table 4: Largest 25 Tracts for College Dorms GQ 500s where over 99 percent of the people in the MAFD match to another GQ

These could possibly be candidates for doing what was just done for the previous patch. Additonal MAFIDs and population were identified by 12/09/20 rerun

Obs	BCUSTATEFP	BCUCOUNTYFP	BCUTRACTCE	_TYPE_	_FREQ_	sum_scount
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	4,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	3,400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	2,700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	2,400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	2,100
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,800
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,100
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	900
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	750
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	350
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	**32,000**

DRB Approval Number: CBDRB-FY21-DSEP-002 Statistics have been rounded according to Census Bureau disclosure standards

Initial Results: Implications of COVID for Counts of Off-Campus Housing Units & Students

Jonathan Spader

July 9, 2020

Pre-Decisional: Internal Use Only

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Introduction

- Research Questions:
  - Do the current RTAD responses include fewer off-campus *housing units* than expected?
  - Do the RTAD responses for off-campus housing units show *smaller household sizes* than expected?

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Data & Research Design

- 2020 Current Responses: RTAD (Real Time Analysis of Data):
  - County-level counts of responses as of 5/31
  - Excludes NPC responses that have been checked in but not processed

- Expected Responses: 2010 Decennial early responses
  - Limit data to housing units in TEA 1 (mail responses)
  - County-level counts of responses as of 4/30 (NRFU began 5/1)
  - Adjust the counts for household growth using ACS 5-year estimates

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Limitations of this Design

Differences between RTAD vs. 2010 Early Responses:

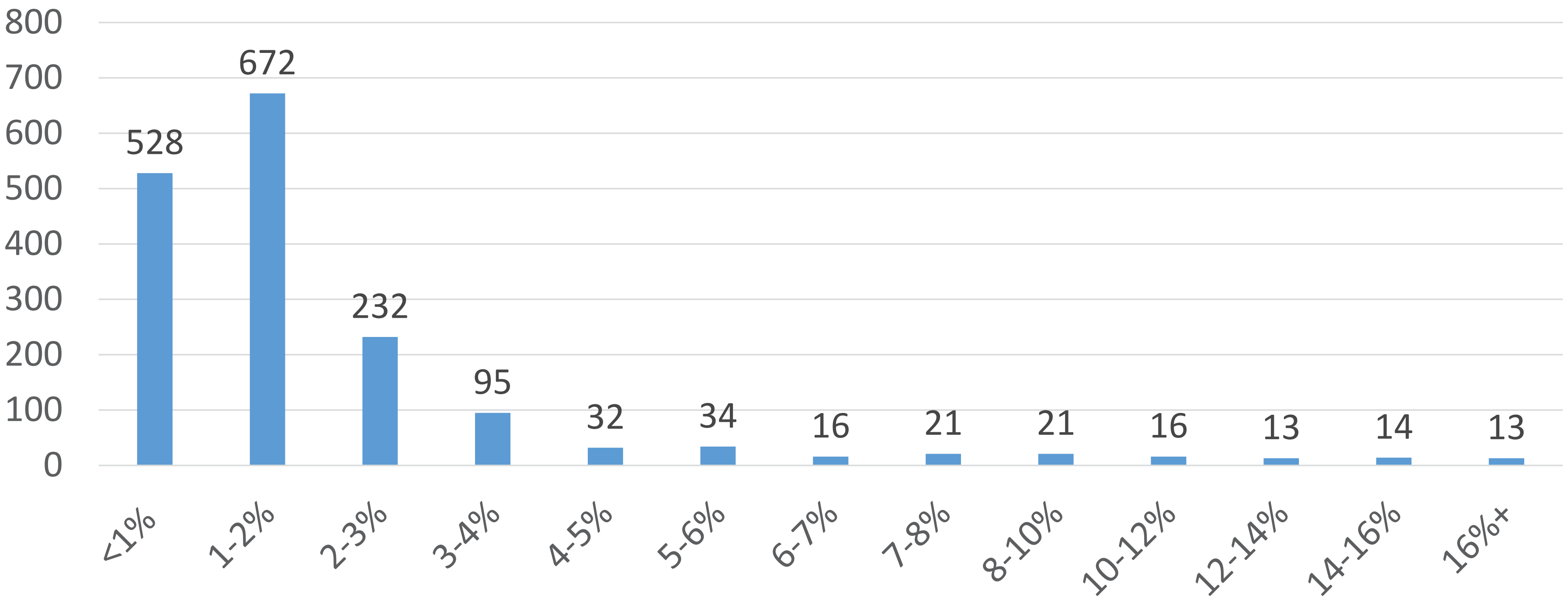
1. Mode: 2020 primarily internet vs. 2010 mail only
2. Timing: 2020 cutoff is 5/31 vs 2010 cutoff is 4/30
3. COVID: Response patterns may be affected through channels other than college closures

Assumption: Any confounders wouldn't disproportionately reduce responses from BOTH (1) college-age householders AND (2) in college areas.

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Number of Counties by % of College-Age Householders

y-axis: # of counties
x-axis: % of households in county headed by an 18-29 year-old college student



% of Households in County with Householder Age 18-29 & Enrolled in Post-Secondary Education

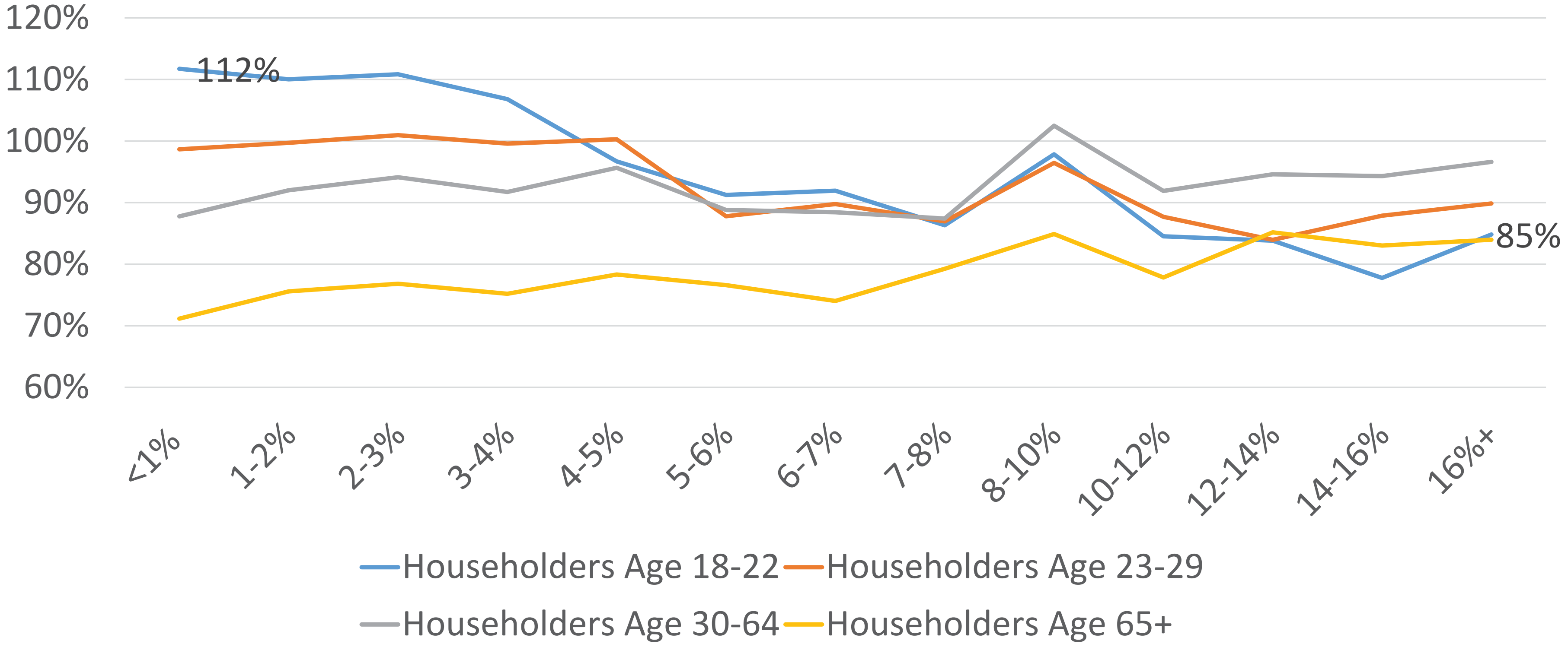
Source: ACS 5-year estimates for 2014-2018.

Pre-Decisional: Internal Use Only

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Ratio of 2020 RTAD Responses / “Expected” Responses

y-axis: Ratio of RTAD responses / expected responses
x-axis: % of households in county headed by an 18-29 year-old college student



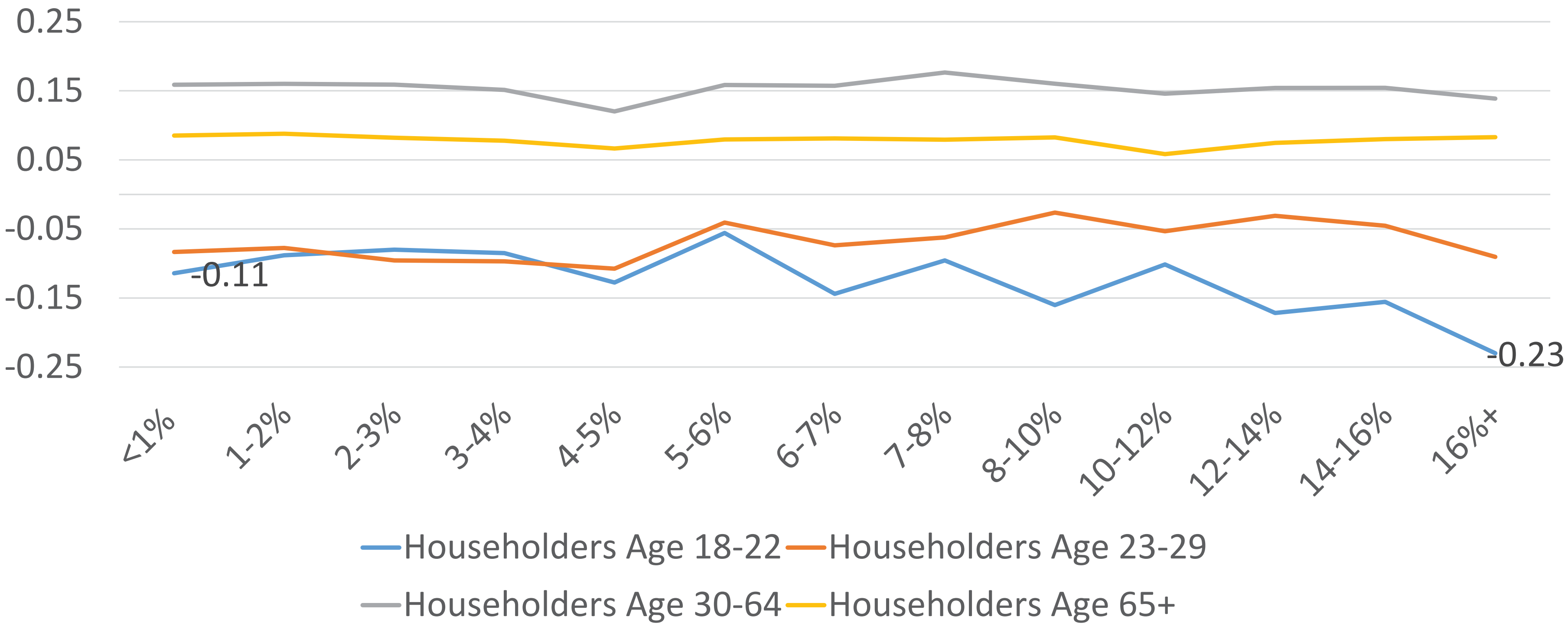
Notes: Expected responses are defined as the # of 2010 mail responses received prior to the start of NRFU (by 4/30/2010) multiplied by a household growth rate estimated from ACS.

Pre-Decisional: Internal Use Only

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Difference in Average Household Size: 2020 minus 2010

y-axis: 2020 RTAD Avg # people per household minus 2010 Final_Pop
x-axis: % of households in county headed by an 18-29 year-old college student



Notes: The 2020 average people per household is calculated from the processed NPC responses. The 2010 final_pop variable is the CEF variable for # people per household.

Pre-Decisional: Internal Use Only

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Next Steps

1. Feedback?
   - Any possible confounders?
   - What additional analyses would be useful?

2. Refine Analysis
   - Tract-level data will be available soon
   - Improve the adjustment for household growth

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Extra Slides

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

10 Counties with Highest % HHs 18-29 & Enrolled in Post-Secondary Education

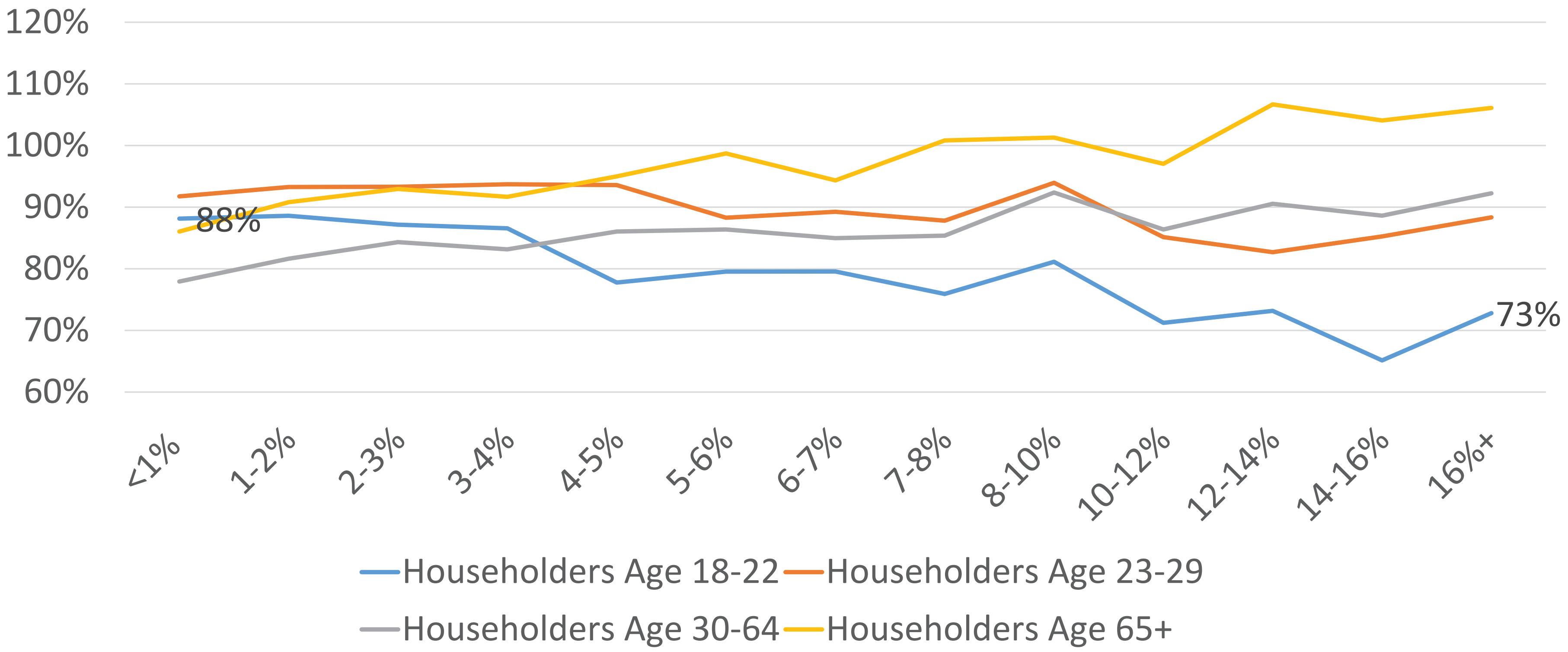
% HHs Age 18-29 College	County	State	University
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]

Pre-Decisional: Internal Use Only

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Alt: Expected Responses = 2010 x County Growth Rate

y-axis: Ratio of RTAD responses / expected responses
x-axis: % of households in county headed by an 18-29 year-old college student



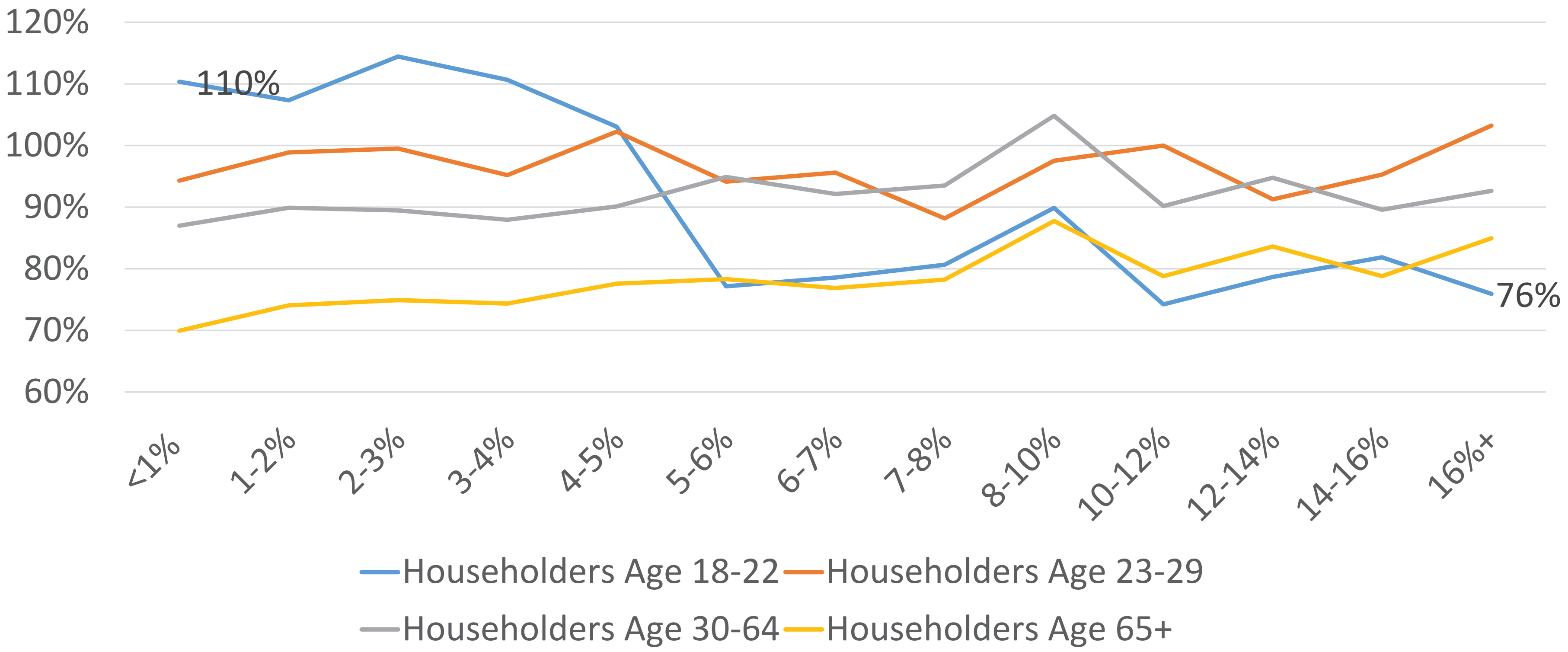
Notes: Expected responses are defined as the # of 2010 mail responses received prior to the start of NRFU (by 4/30/2010) multiplied by a household growth rate estimated from ACS.

Pre-Decisional: Internal Use Only

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Alt: Estimate Growth using County, Age, & College Share

y-axis: Ratio of RTAD responses / expected responses
x-axis: % of households in county headed by an 18-29 year-old college student



Notes: Expected responses are defined as the # of 2010 mail responses received prior to the start of NRFU (by 4/30/2010) multiplied by a household growth rate estimated from ACS.

Pre-Decisional: Internal Use Only

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Sample

- 1,707 counties in the analysis sample
  - Includes 54% of counties and 94% of all occupied units in the U.S.
  - Excludes counties with few housing units in each age group due to concerns about outlier values

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

EXHIBIT 5

May 27 Correspondence Granting Expedited Processing

From: admin@foiaonline.gov
Sent: Friday, May 28, 2021 1:03 PM
To: Jason Torchinsky
Subject: FOIA Expedited Processing Disposition Reached for DOC-CEN-2021-001311

Your request for Expedited Processing for the FOIA request DOC-CEN-2021-001311 has been granted. Additional details for this request are as follows:

- Request Created on: 04/07/2021
- Request Description: All summaries, “tabulations[,] and other statistical materials,” 13 U.S.C. § 8(b), derived from, summarizing, and/or otherwise relating to the original underlying group quarters population data for Census Day, April 1, 2020, received in response to the Census Bureau’s 2020 Group Quarters Enumeration questionnaire regarding institutional living facilities or other housing facilities. In requesting these summaries, “tabulations[,] and other statistical materials,” we do not seek disclosure of the underlying raw group quarters population data itself as originally “reported by, or on behalf of, any particular respondent” to the Bureau, 13 U.S.C. § 8(b), nor do we seek any “publication whereby the data furnished by any particular establishment or individual under this title can be identified,” 13 U.S.C. § 9(a)(2); instead, we seek records deriving from or summarizing the originally reported raw data, and/or records with data that has been reformulated or repurposed by the Bureau in a form such that the underlying data can no longer be identified with a particular establishment or individual. For instance, any statewide aggregate total group quarters population tabulations of data that exclude, omit, or redact the original group quarters numbers as reported by, or on behalf of, individual institutions (i.e., tabulations where the Bureau excluded the underlying individualized raw data, or where such data can be redacted from the tabulations while producing the aggregate population totals) would be responsive to this request. Please note that this request encompasses both digital and physical records. “Record” should be understood as that term is defined under FOIA (5 U.S.C. § 552(f)(2)), and applicable case law (see, e.g., Forsham v. Harris, 455 U.S. 169, 193 (1980)), existing in any format whatsoever. Please understand “Census Bureau” to include any employees working for the Bureau...
- Expedited Processing Original Justification: Fair Lines requests that the processing of this request be expedited pursuant to 15 C.F.R. § 4.6(f).
  This request qualifies for expedited processing both because it involves “[a] matter of widespread and exceptional media interest involving questions about the Government’s integrity which affect public confidence.” 15 C.F.R. § 4.6(f)(iii). Indeed, there are few matters of more widespread interest than the integrity of our election system and democracy; issues regarding the accuracy and collection of group quarters data and its potentially significant impact on the redistricting process for states are integrally connected to these critical matters...
- Expedited Processing Disposition Reason: N/A

UNITED STATES DEPARTMENT OF COMMERCE
U.S. Census Bureau
Washington, DC 20233-0001

May 28, 2021

Mr. Jason Torchinsky
Fair Lines American Foundation, Inc.
2308 Mount Vernon Ave., Suite 716
Alexandria, VA 22301
jtorchinsky@hvjt.law

Dear Mr. Torchinsky:

This letter is in further response to your Freedom of Information Act (FOIA), Title 5, United States Code, Section 552, request dated March 31, 2021, to the U.S. Census Bureau's FOIA Office. We received your request in this office on April 7, 2021, assigned it tracking number DOC-CEN-2021-001311, and are responding under the FOIA to your request for all summaries, tabulations, and other statistical materials derived from, summarizing, and/or otherwise relating to the original underlying group quarters population data for Census Day, April 1, 2020, received in response to the Census Bureau's 2020 Group Quarters Enumeration questionnaire regarding institutional living facilities or other housing facilities.

In your request correspondence, you seek a waiver of fees. After review of your request, we have determined that your fee waiver justification is sufficient to grant your request for a waiver of processing fees. Therefore, in accordance with 15 CFR Section 4.11, we are granting your request for a fee waiver.

In your request correspondence, you also seek expedited processing of your request. After review of your request, we have determined that your expedited processing justification is sufficient to grant your request for expedited processing. Therefore, in accordance with 15 CFR Section 4.6, we are granting your request for expedited processing. Your FOIA request will be placed on the priority processing track and processed as soon as practicable.

We appreciate your interest in the Census. Because this matter is now in litigation, if you have any further questions concerning this letter, please contact Jonathan Kossak at Jonathan.Kossak@usdoj.gov.

Sincerely,

Vernon Curry

Vernon E. Curry, PMP, CIPP/G
Freedom of Information Act/Privacy Act Officer
Chief, Freedom of Information Act Office

EXHIBIT 6

July 6 Additional Production

On Jul 6, 2021, at 9:32 PM, Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov> wrote:

Dear Jason and Ken,

Thanks for your email last week. As you note, Defendants redacted only 115 of the 988 pages they produced on May 24, 2021, in response to your client's FOIA request in this litigation. As I explained in my June 25, 2021 email, the redactions were made by the Census Bureau's Disclosure Review Board (DRB), whose purpose is to support the Data Stewardship Executive Policy (DSEP) Committee to ensure that every information product released by the Census Bureau adheres to the confidentiality requirements of Title 13 and other applicable statutes. As you are aware from the face of your client's request, 13 U.S.C. §§ 8(b) and 9 are the statutory provisions under the Census Act that impose a mandate upon the Census Bureau to protect the confidentiality of individual census responses and data. These provisions prohibit the Census Bureau from releasing "any publication whereby the data furnished by any particular establishment or individual under this title can be identified," and allows the Secretary to provide aggregate statistics so long as those data "do not disclose the information reported by, or on behalf of, any particular respondent."

Other than the inconsistency you purport to identify in the first bullet of your email, the remainder of your concerns appear to be driven by your misconception of how the Title 13 confidentiality provisions work. You contend that the DRB improperly redacted certain data because "it is only derived from raw data, but does not include the numbers that were furnished by any particular establishment or individual"; or that certain "statistical information or tabulations . . . do not disclose any raw data reported by particular respondents"; or that certain "categories of data described are clearly summary in nature, and would not lead to disclosure of any particular respondent's reported data." These arguments, and those repeated in the same or similar wording in your other bullets, are all based upon the same erroneous conception of Title 13's confidentiality provisions.

As you are aware from the State of Alabama litigation in which you participated, to satisfy Title 13's privacy strictures, the Census Bureau must account for "complementary disclosure," which is the release of data that does not appear to contain individually identifiable information, but could result in identifying individuals when those data are coupled with other information in existing Census Bureau publications or other publicly available information. As you are also aware from the Alabama case, the Census Bureau has dedicated significant resources to addressing the Fundamental Law of Information Reconstruction, which says that overly accurate estimates of too many statistics can destroy privacy. Modern computational and information resources feed on statistical data, and the cumulative effect of statistical releases in this age of computing power and sophistication poses a significant threat to the privacy of individual responses. The Census Bureau generally avoids the release of intermediate work product because it can be used in combination with other intermediate work products, official publications, and the final product to re-identify individual respondents and their data items.

The DRB reviewed the 988 pages produced to you and determined that the withheld data had to be redacted because its release would violate Title 13's confidentiality provisions in light of complementary disclosure and/or reconstruction concerns. I know of no FOIA case (nor any other case in any other context) that undermines the Census Bureau's authority to redact this information. Indeed, the last significant challenge in the context of FOIA to the Census Bureau's withholding of information pursuant to Title 13's confidentiality provisions was *Baldridge v. Shapiro*, 455 U.S. 345 (1982), in which the Supreme Court reviewed the history of those provisions and determined that Congress's intention in establishing the confidentiality provisions, "was to encourage public participation and maintain public confidence that information given to the Census Bureau would not be disclosed." *Id*. at 361. *Baldridge* is nearly 40 years old and technology has greatly advanced since then. The Census Bureau has to keep up with the technology to maintain the public's confidence. Title 13's confidentiality provisions would be severely undermined if the Census Bureau did not take into account the risk of re-identification attacks on aggregated data releases. Accordingly, the redactions you identify below are not "improper." We are confident they will stand against challenge in any court.

However, as I mentioned on our last call, any such challenge is premature. Motions for partial summary judgment in FOIA cases are heavily disfavored by the courts in this jurisdiction, and you have not identified any particular reason why the redacted data is needed urgently. Moreover, you already have received the vast majority of information in an unredacted manner, and the Census Bureau will be publicly releasing vast quantities of data no later than August 16, 2021. Your client has asked for emails responsive to its FOIA request, and Defendants have identified 917 potentially responsive emails, consisting of 25,899 pages of material. That does not include either attachments to those emails or Excel spreadsheets. The attachments increase the number of documents to 2,414 and the page count to 35,880 pages. The Excel spreadsheets, which would be produced in native format, have to be converted into pdfs to get a page count. The total page count figure for the excel spreadsheets would be 760,000. That is obviously an astronomical figure. In the ordinary course of a FOIA litigation, we would work with a plaintiff to figure out how to narrow the universe of potentially responsive material down to reasonable proportions, but that takes time. As stated, Defendants will use their best efforts to process 300 pages of potentially responsive records every month. It may be that in 2-4 months your client determines that "the juice is not worth the squeeze," and agrees to forego further processing. Or your client may identify certain materials in the disclosed records that it finds useful and may agree to narrow the universe of material to be reviewed. We are happy to continue negotiating the parameters of your request, but such negotiations are likely to be more productive after a few months of processing have taken place.

Given the early stage of this litigation, we intend to oppose as premature any motion for partial summary judgment you seek leave to file. And even if the Court allows it, we will move to stay the processing of any additional records until after the briefing process is complete, since that process will take up the resources of key staff who would otherwise be participating in the processing of potentially responsive records.

Finally, attached are the two additional “post-December 2020” documents we have been discussing in the emails below and in our last call. As for your concern that it seems unlikely that there are only two such documents, the Census Bureau has verified for us that the documents produced are the only ones responsive to your FOIA request. For your awareness, Defendants have employed the typical “date-of-search” temporal limitation blessed by the D.C. Circuit. For the post-December 2020 records, the date the search for those records began was May 19, 2021.

I’m happy to discuss any of the above in more depth this week. Please let me know when you are available.

- Jonathan

GQ and DRF1 Review Update

Marc Perry, POP
Christine Borman, POP
Jonathan Spader, SEHSD

December 31, 2020

Shape
your future
START HERE >

United States®
Census
2020

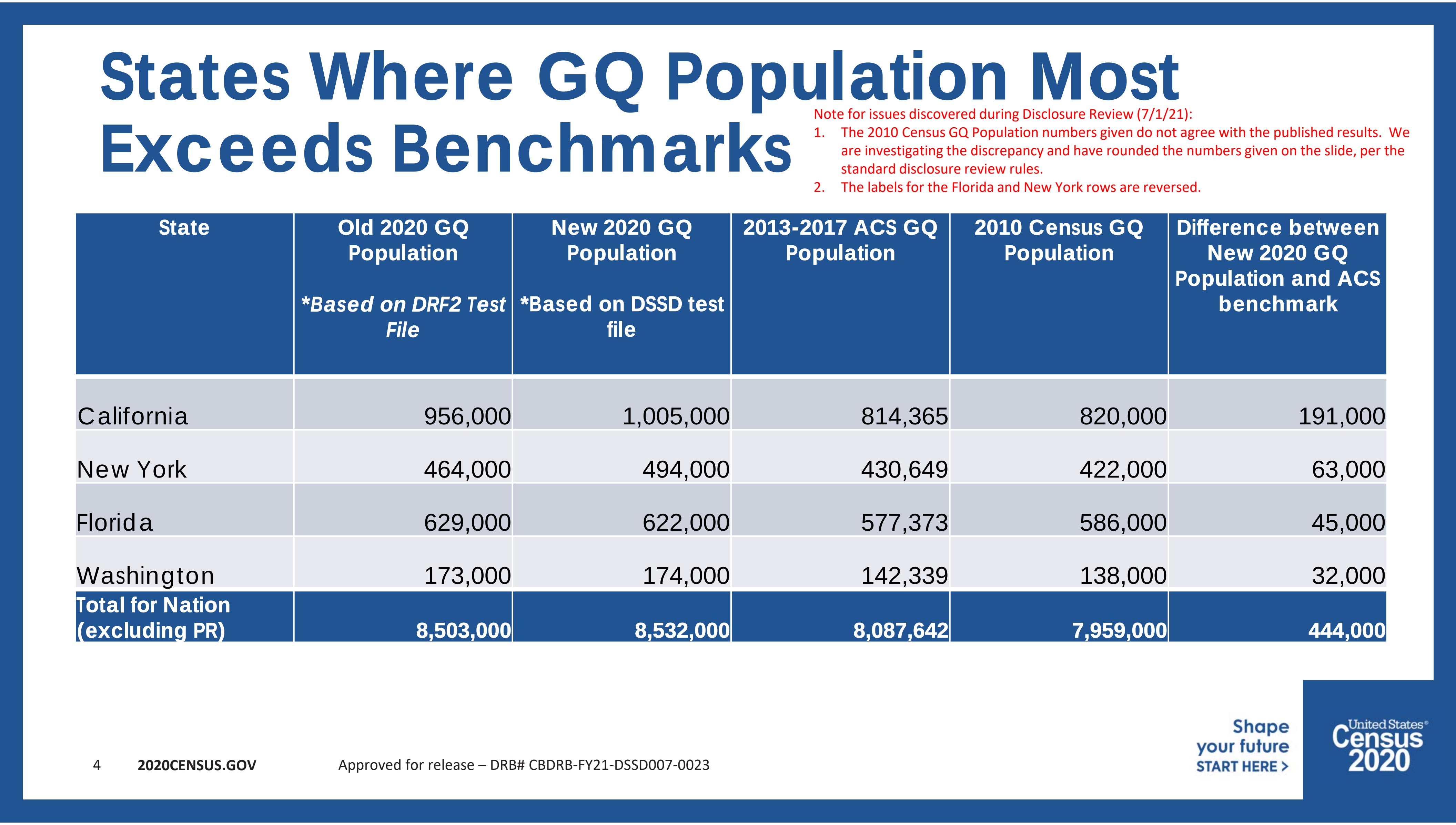
Agenda

Topics:

- Group Quarter Updates
- DRF1 Item Nonresponse Rates

Group Quarters (GQ) Recap

- Review of the DRF1 uncovered numerous anomalies in the GQ data. In some instances, GQ populations were either zero or well below benchmark data. In other cases, facility-level or census tract data appeared too high, and manual inspection revealed erroneous duplications of GQ unit level data.
- DSSD then performed an unduplication and imputation exercise, in addition to the cross-directorate effort to contact thousands of GQs to get population counts.
- The new data set from DSSD that contains unduplication and imputation does give more reasonable GQ counts for many GQ units that appeared to have been undercounted or not enumerated.
- But sometimes the new data set artificially inflates the GQ populations, making them larger than benchmark estimates. Consequently, the U.S. GQ population is nearly 8.6 million, which exceeds the benchmark estimate of about 8.1 million. California is considerably above benchmark estimates.



States Where GQ Population Most Exceeds Benchmarks

Note for issues discovered during Disclosure Review (7/1/21):

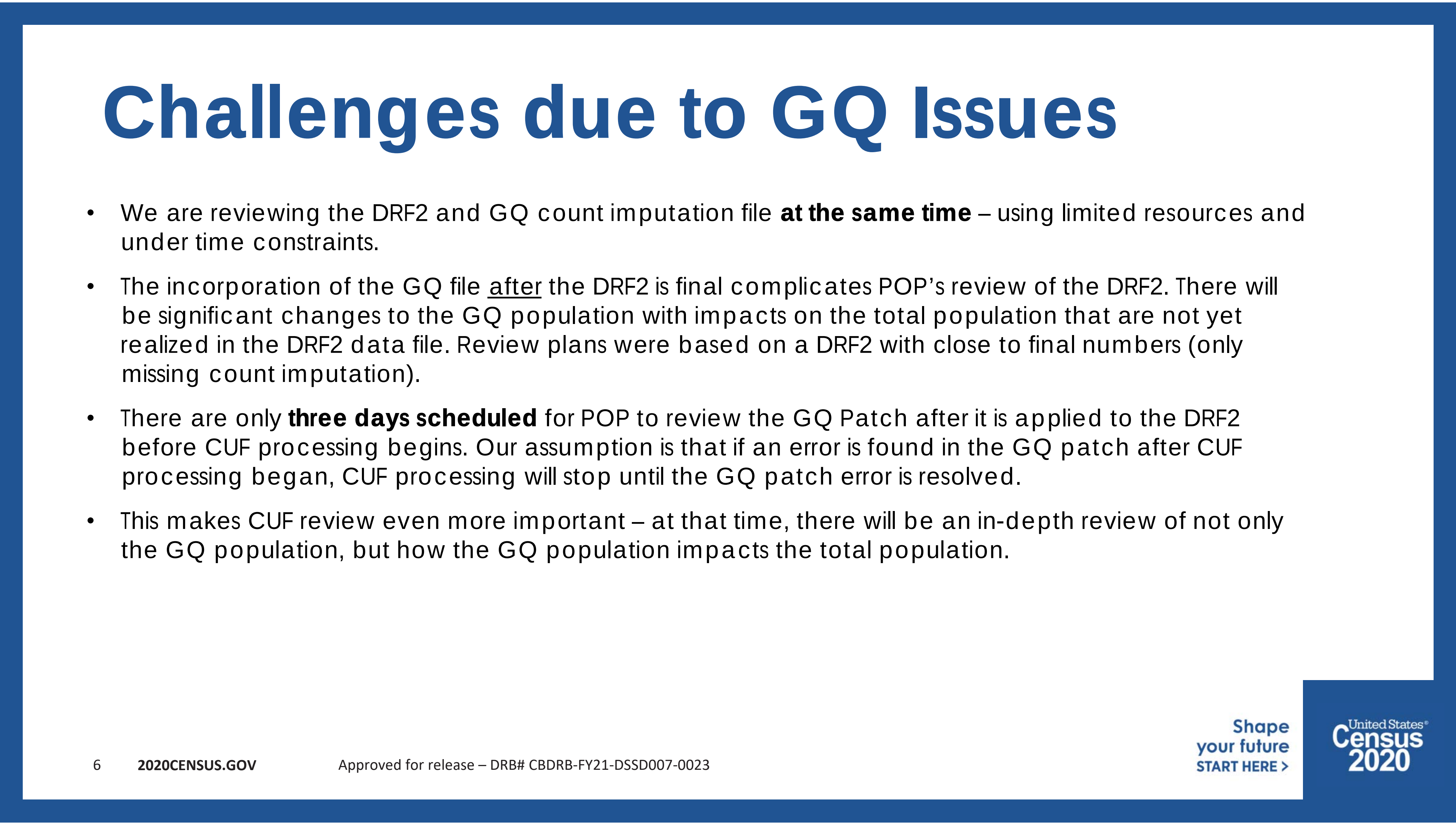
1. The 2010 Census GQ Population numbers given do not agree with the published results. We are investigating the discrepancy and have rounded the numbers given on the slide, per the standard disclosure review rules.
2. The labels for the Florida and New York rows are reversed.

State	Old 2020 GQ Population **Based on DRF2 Test File*	New 2020 GQ Population *Based on DSSD test file	2013-2017 ACS GQ Population	2010 Census GQ Population	Difference between New 2020 GQ Population and ACS benchmark
California	956,000	1,005,000	814,365	820,000	191,000
New York	464,000	494,000	430,649	422,000	63,000
Florida	629,000	622,000	577,373	586,000	45,000
Washington	173,000	174,000	142,339	138,000	32,000
Total for Nation (excluding PR)	**8,503,000**	**8,532,000**	**8,087,642**	**7,959,000**	**444,000**

Shape your future START HERE >

Review of GQ Files

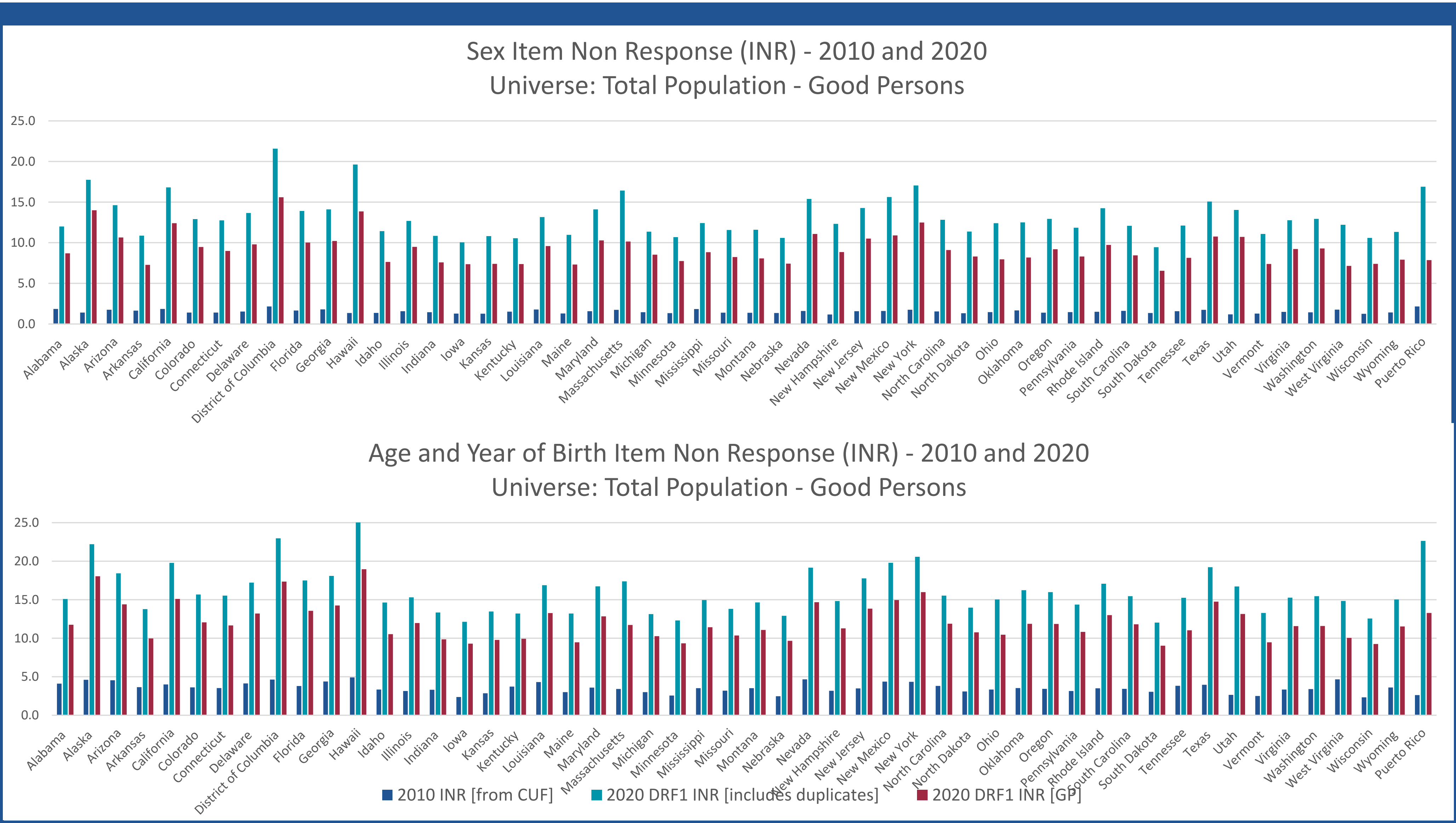
- Manual inspection and adjudication of the population counts for each GQ unit is complicated and requires the evaluation of the impact of multiple processes (unduplication, imputation and new data sources from call backs). It is not feasible to evaluate all GQ populations at the unit level and it is not possible to evaluate all GQ populations at the facility level.
- One approach would be to use the new DSSD data set (that includes unduplication as well as imputation of some GQ types), and to evaluate the resulting county level GQ populations against benchmark data to identify instances where the results do not appear to be reasonable.
  - In **33 counties** the 2020 Census GQ population is **greater than the benchmark** by 5,000 or more
  - In **7 counties** the 2020 Census GQ population is **below the benchmark** by at least 5,000.
  - In those 40 counties, we could do a manual inspection of the GQ unit-level data to better ascertain the cause of the anomalous count and determine next steps. In many instances where the GQ count exceeds benchmark estimates and there appears to be erroneous duplication of units, we would recommend turning off imputation for the duplicate units. In other instances, we may suggest not taking the results from unduplication or call-back information (for example, when the call-back number is for the entire facility). Modifications to the methodology itself may also be necessary.



Challenges due to GQ Issues

- We are reviewing the DRF2 and GQ count imputation file **at the same time** – using limited resources and under time constraints.
- The incorporation of the GQ file <u>after</u> the DRF2 is final complicates POP's review of the DRF2. There will be significant changes to the GQ population with impacts on the total population that are not yet realized in the DRF2 data file. Review plans were based on a DRF2 with close to final numbers (only missing count imputation).
- There are only **three days scheduled** for POP to review the GQ Patch after it is applied to the DRF2 before CUF processing begins. Our assumption is that if an error is found in the GQ patch after CUF processing began, CUF processing will stop until the GQ patch error is resolved.
- This makes CUF review even more important – at that time, there will be an in-depth review of not only the GQ population, but how the GQ population impacts the total population.

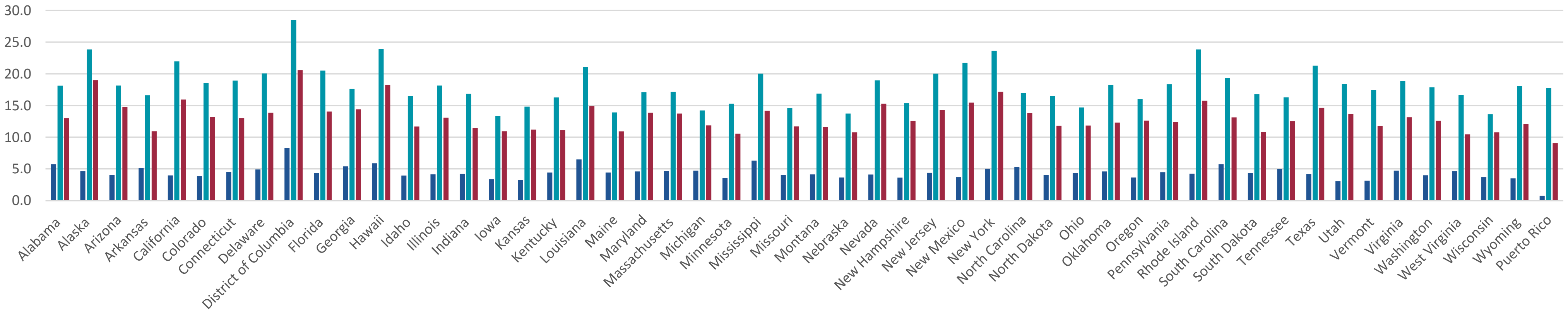
DRF1 Item Nonresponse



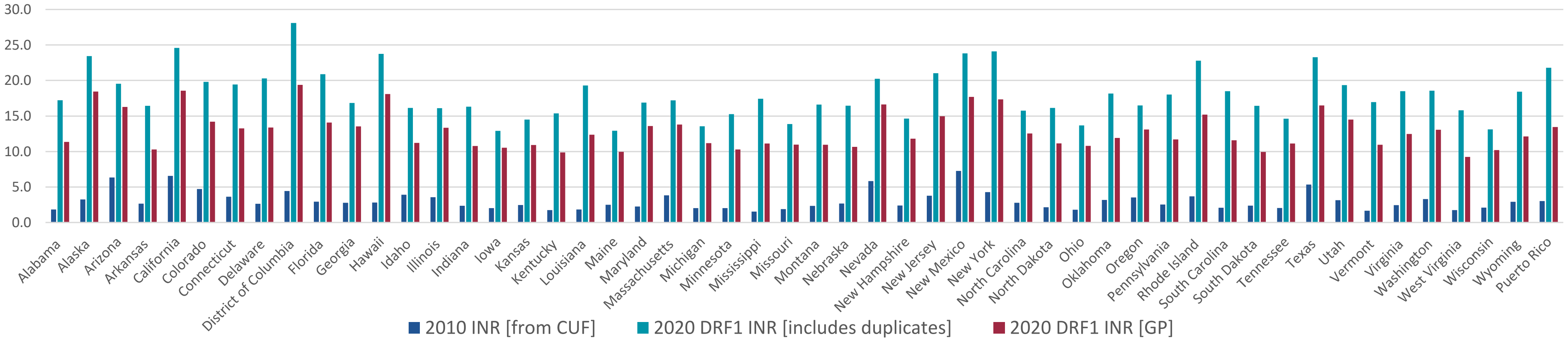
Sex Item Non Response (INR) - 2010 and 2020
Universe: Total Population - Good Persons

Age and Year of Birth Item Non Response (INR) - 2010 and 2020
Universe: Total Population - Good Persons

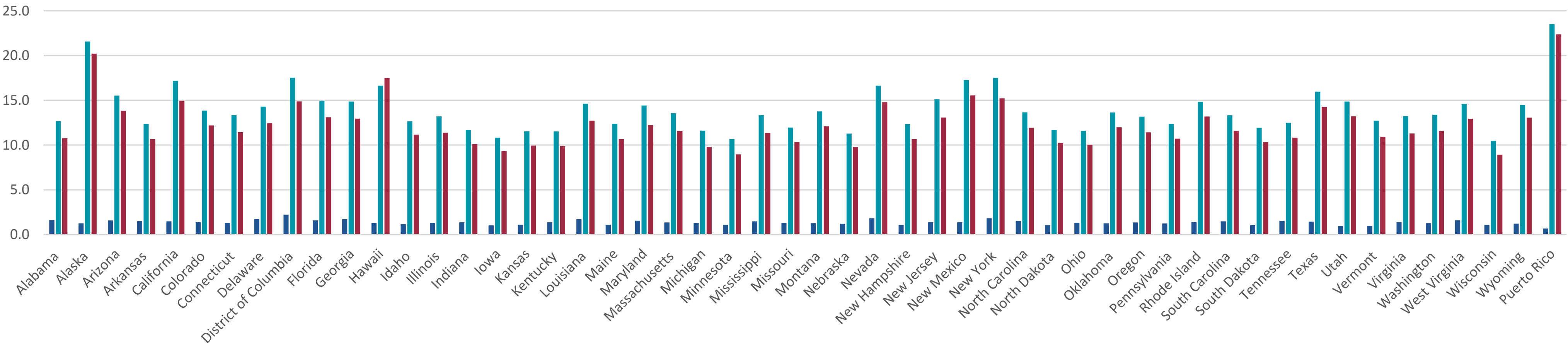
Hispanic Origin Item Non Response (INR) - 2010 and 2020
Universe: Total Population - Good Persons



Race Item Non Response (INR) - 2010 and 2020
Univere: Total Population - Good Persons

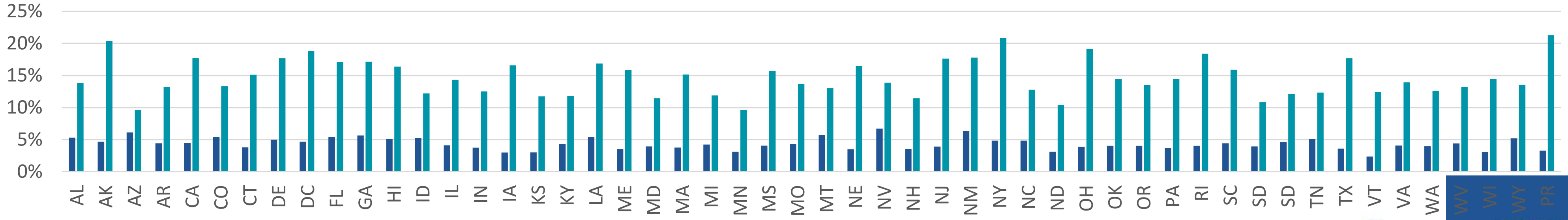


Relationship Item Non Response (INR) - 2010 and 2020
Univere: Household Population - Good Persons



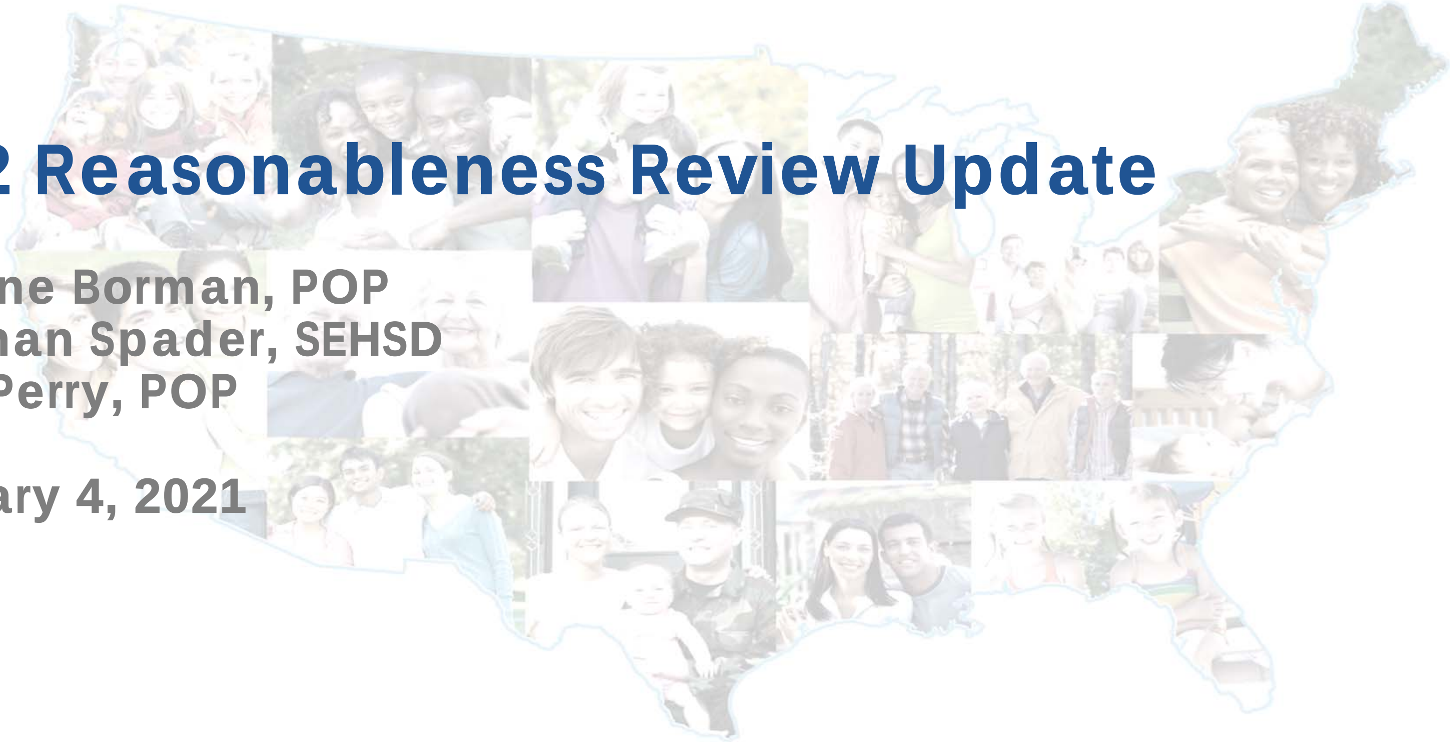
■ 2010 INR [from CUF] ■ 2020 DRF1 INR [includes duplicates] ■ 2020 DRF1 INR [GP]

Tenure Item Non Response (INR) – 2010 and 2020
Universe: Housing Units



■ 2010 INR [from CUF] ■ 2020 DRF1 INR [includes duplicates]

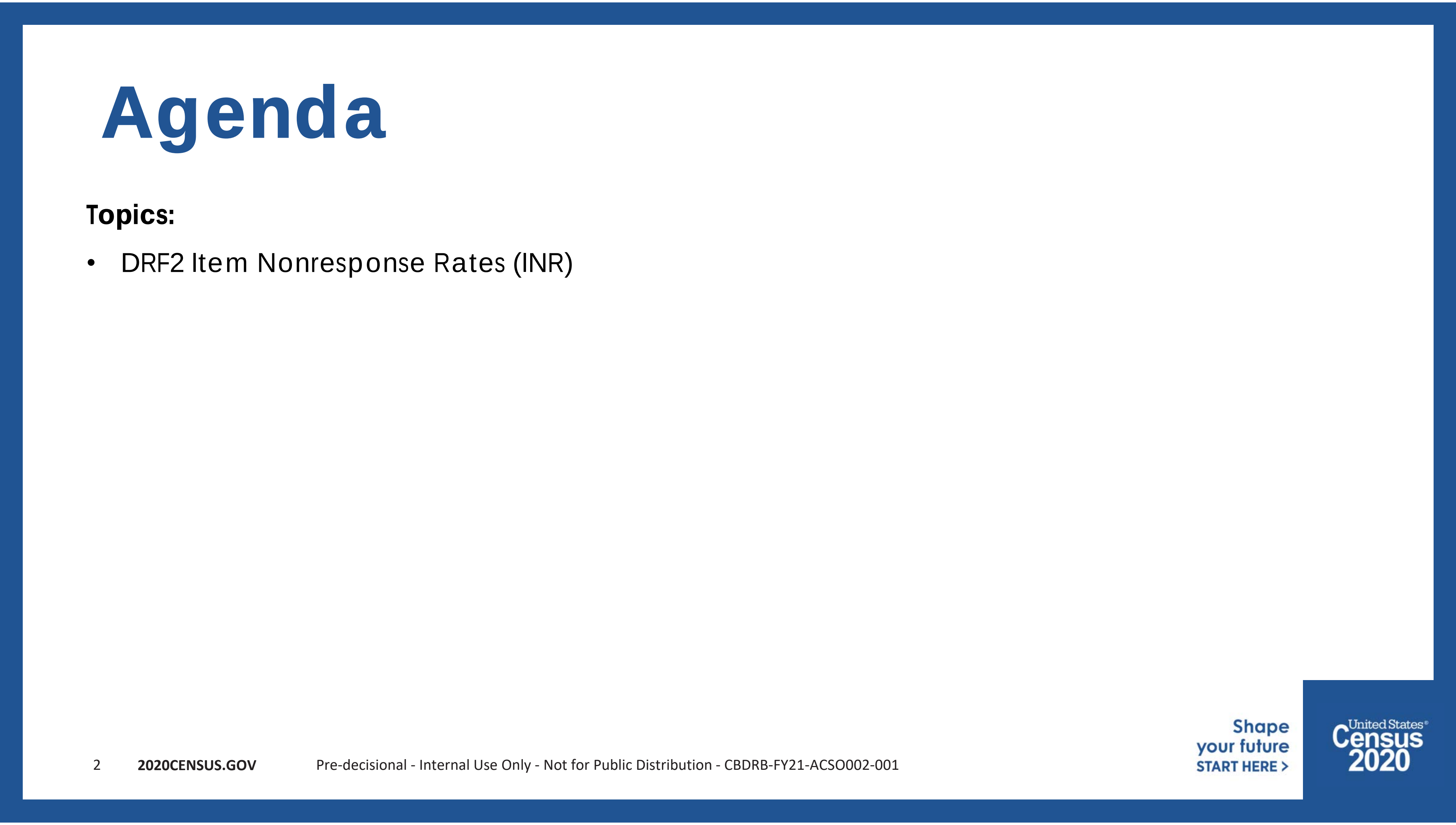
Approved for release – DRB# CBDRB-FY21-DSSD007-0023



DRF2 Reasonableness Review Update

Christine Borman, POP
Jonathan Spader, SEHSD
Marc Perry, POP

February 4, 2021



Agenda

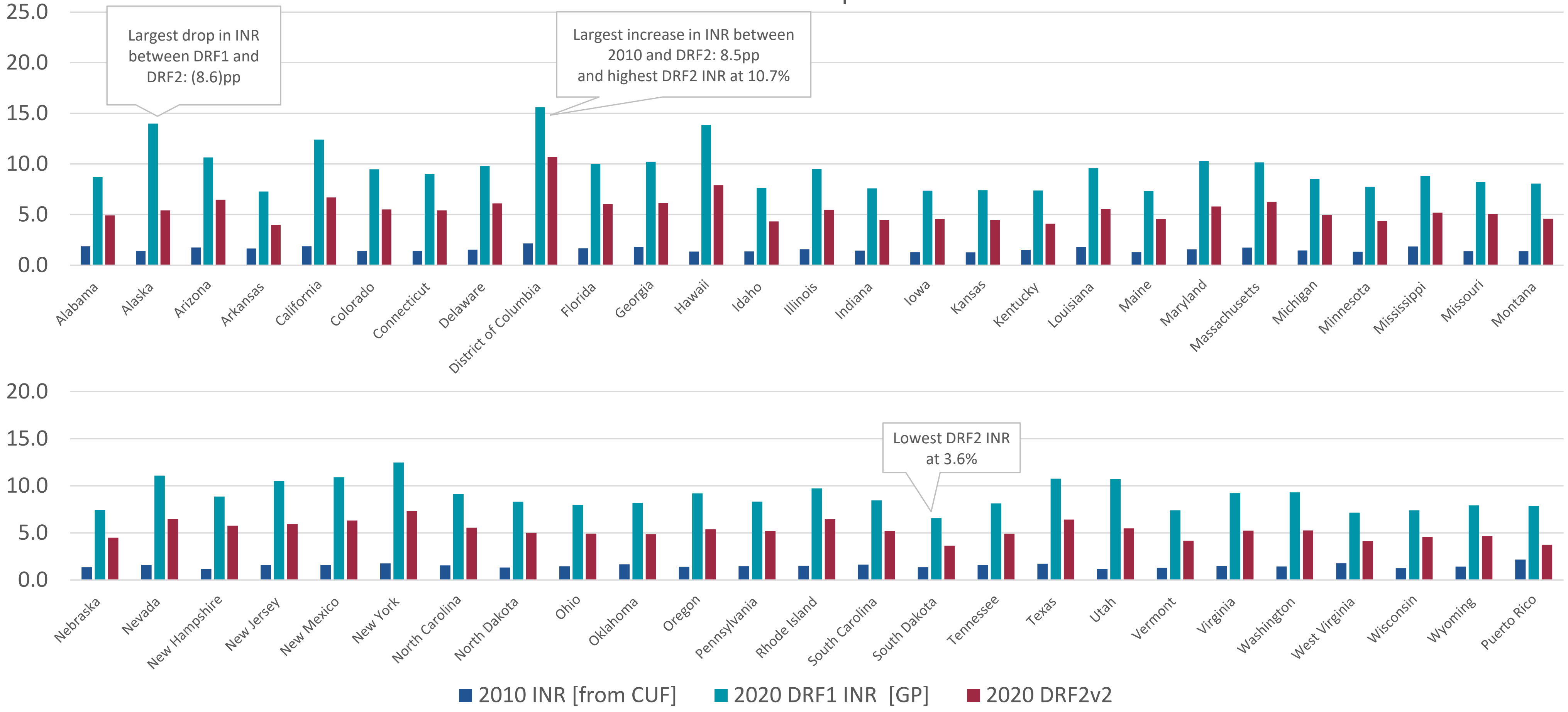
Topics:

- DRF2 Item Nonresponse Rates (INR)

Shape
your future
START HERE >
United States®
Census
2020

Sex Item Non Response - 2010 Census, 2020 DRF1 Good Persons, and 2020 DRF2 Selected Persons

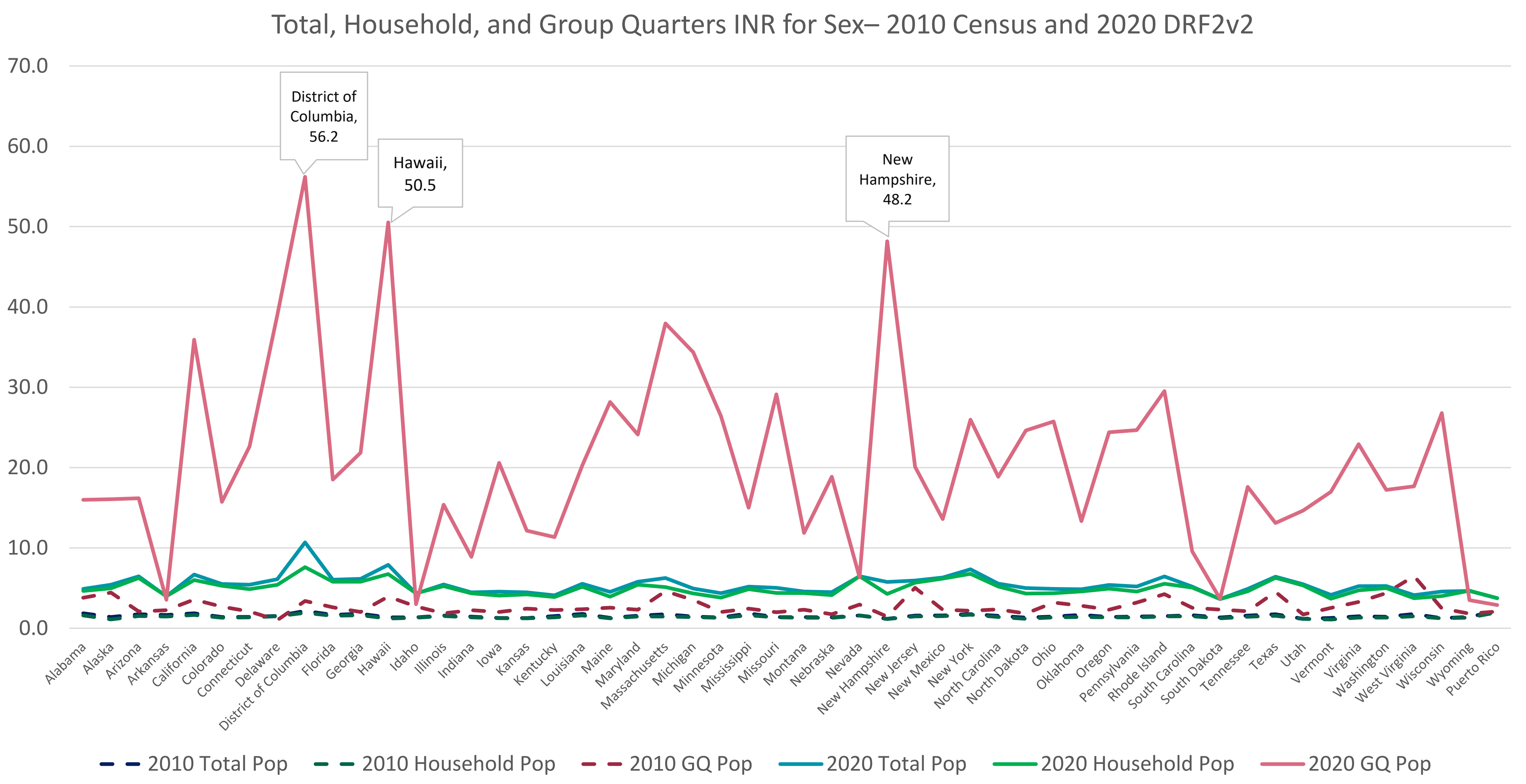
Universe: Total Population



Largest drop in INR between DRF1 and DRF2: (8.6)pp

Largest increase in INR between 2010 and DRF2: 8.5pp and highest DRF2 INR at 10.7%

Lowest DRF2 INR at 3.6%

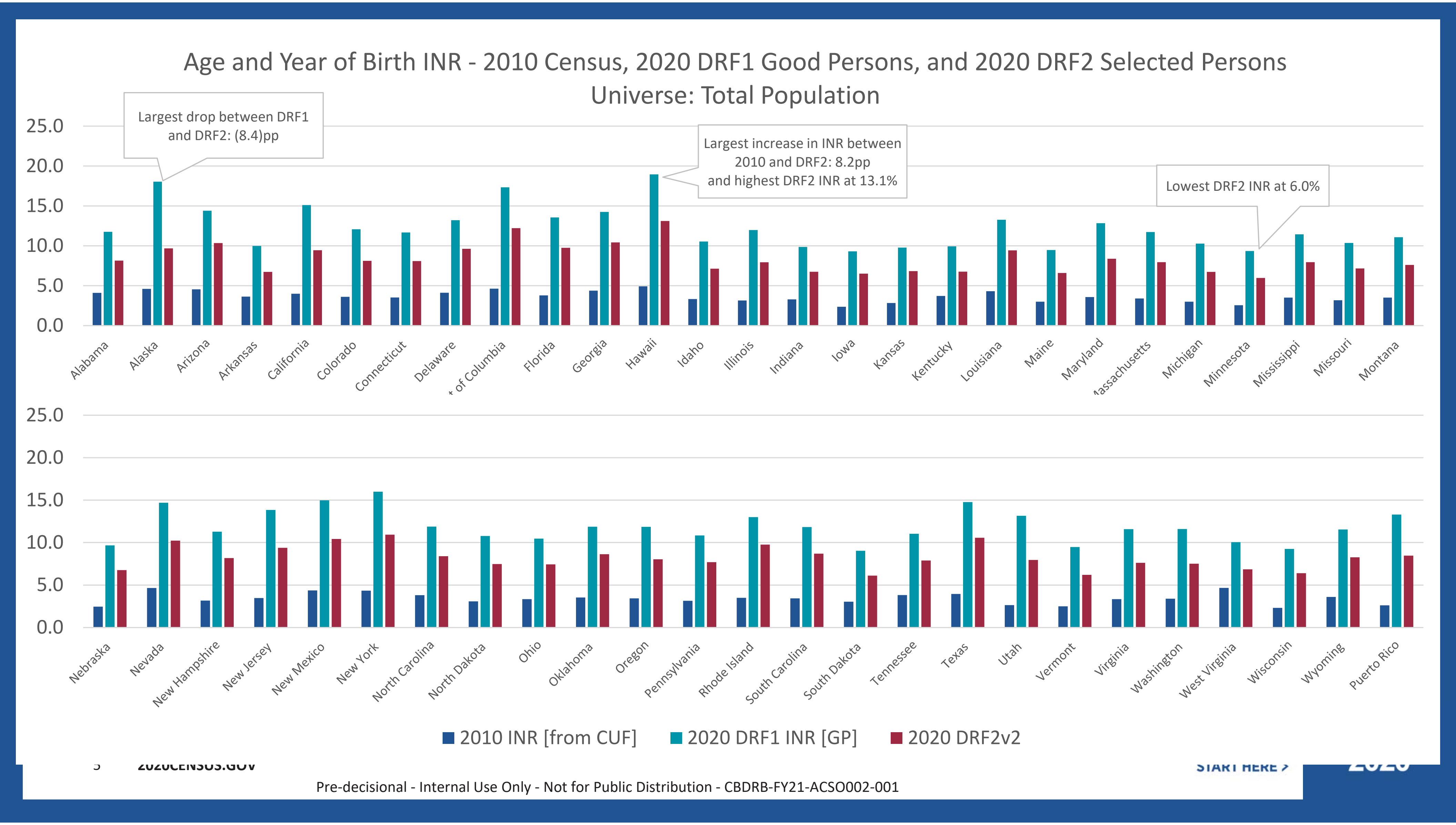


Total, Household, and Group Quarters INR for Sex– 2010 Census and 2020 DRF2v2

District of Columbia, 56.2

Hawaii, 50.5

New Hampshire, 48.2



Age and Year of Birth INR - 2010 Census, 2020 DRF1 Good Persons, and 2020 DRF2 Selected Persons
Universe: Total Population

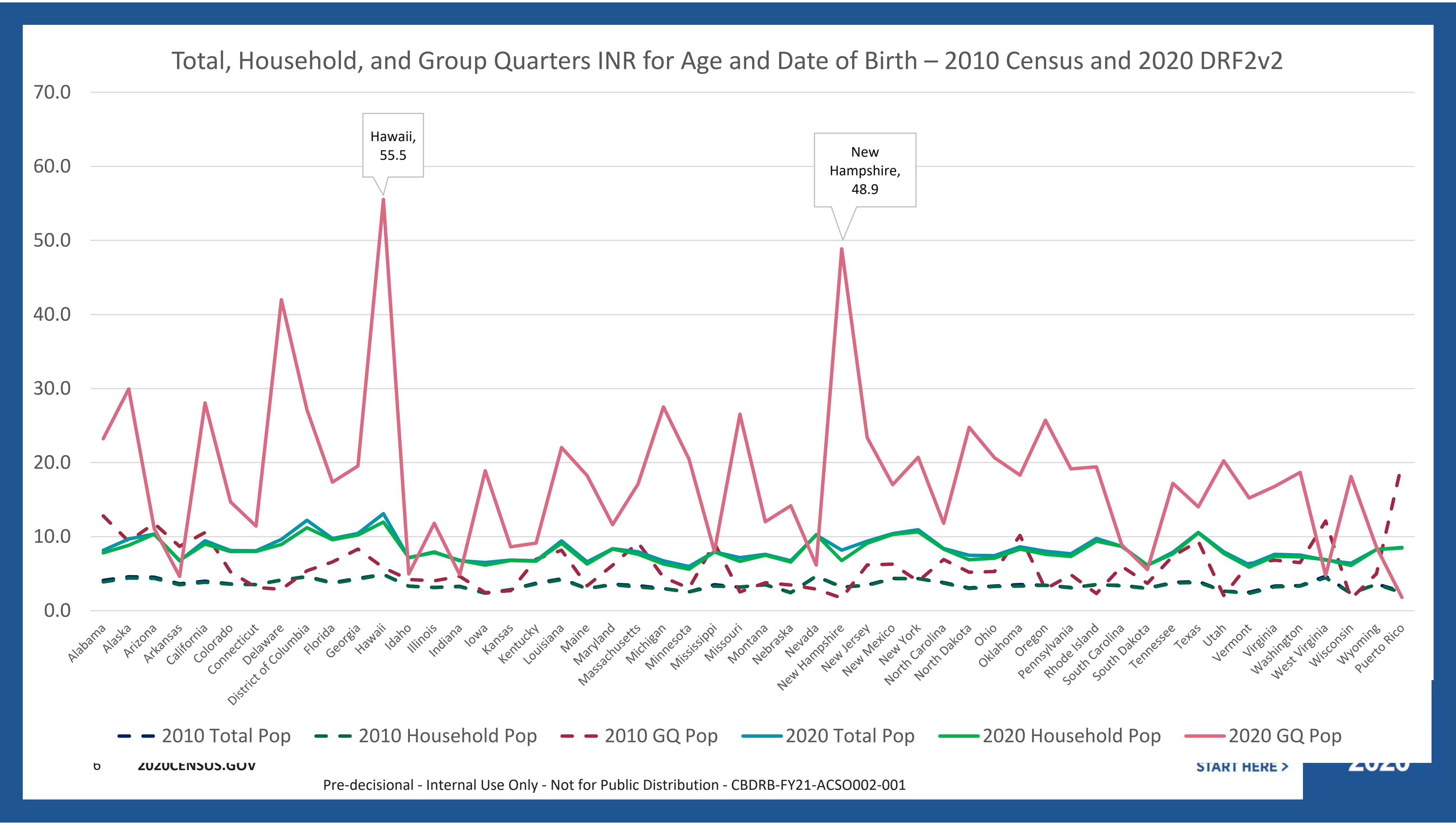
Largest drop between DRF1 and DRF2: (8.4)pp

Largest increase in INR between 2010 and DRF2: 8.2pp and highest DRF2 INR at 13.1%

Lowest DRF2 INR at 6.0%

🟦 2010 INR [from CUF] 🟩 2020 DRF1 INR [GP] 🟥 2020 DRF2v2

Pre-decisional - Internal Use Only - Not for Public Distribution - CBDRB-FY21-ACSO002-001



Total, Household, and Group Quarters INR for Age and Date of Birth – 2010 Census and 2020 DRF2v2

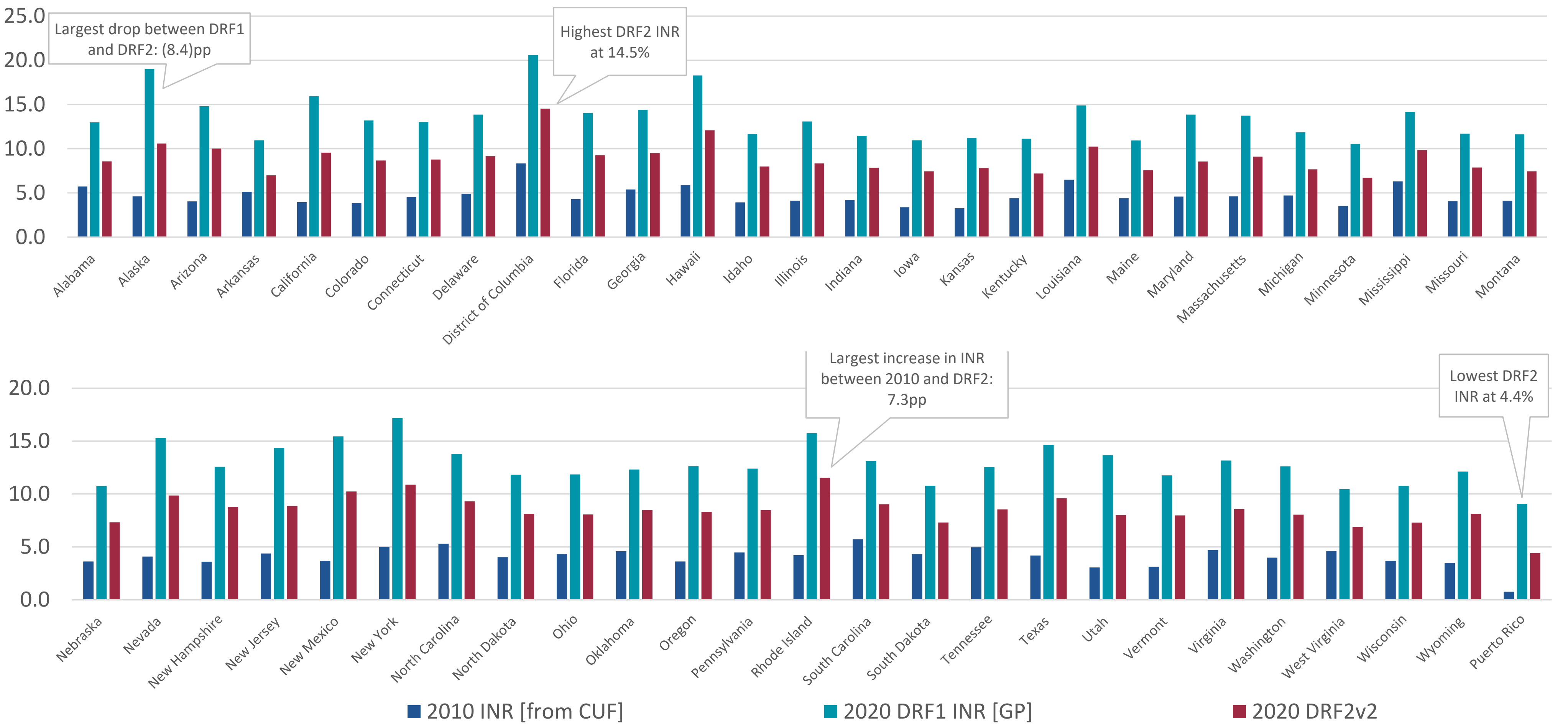
6 2020CENSUS.GOV

START HERE >

2020

Pre-decisional - Internal Use Only - Not for Public Distribution - CBDRB-FY21-ACSO002-001

Hispanic Origin INR - 2010 Census, 2020 DRF1 Good Persons, and 2020 DRF2 Selected Persons
Universe: Total Population



Largest drop between DRF1 and DRF2: (8.4)pp

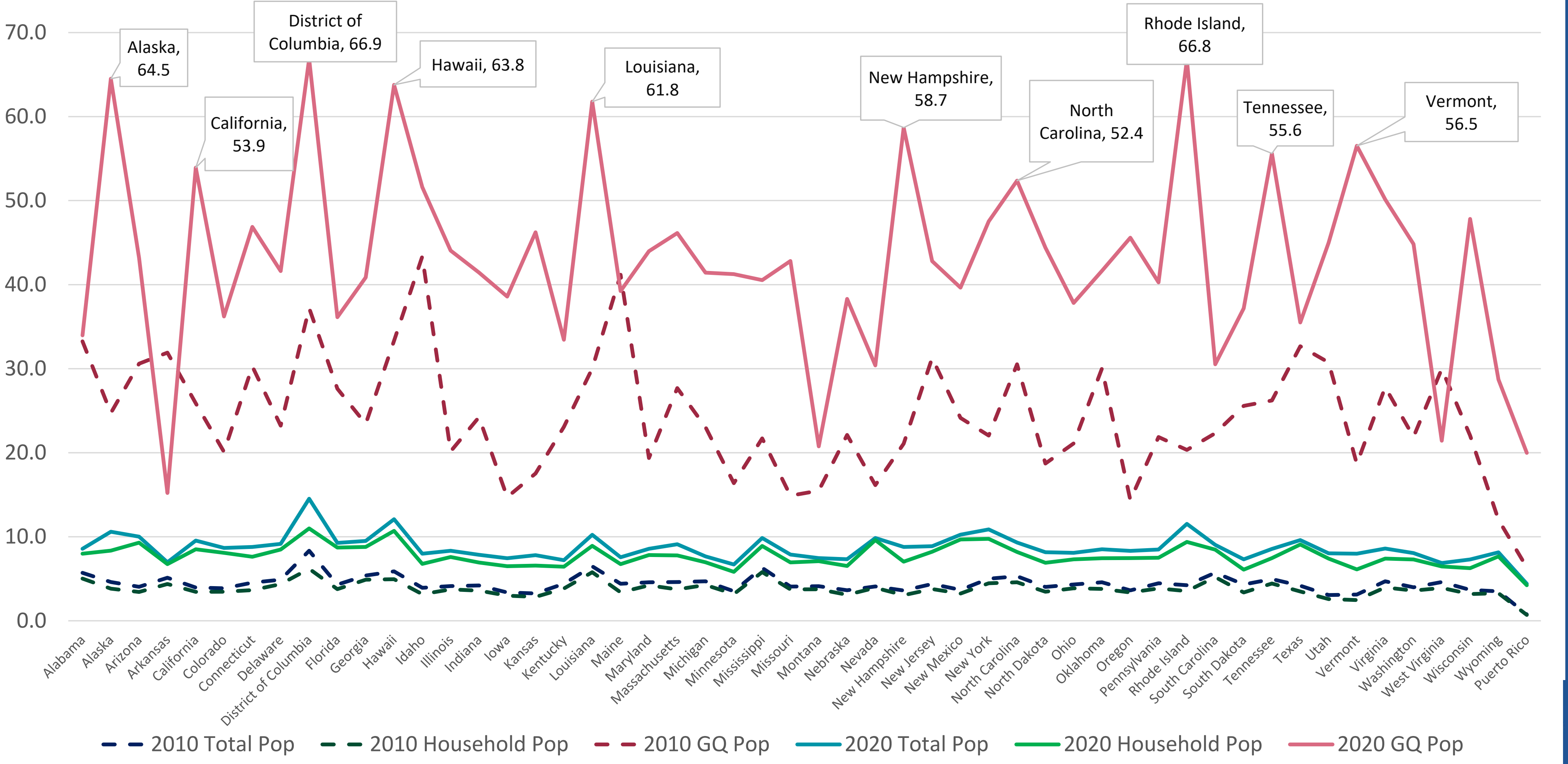
Highest DRF2 INR at 14.5%

Largest increase in INR between 2010 and DRF2: 7.3pp

Lowest DRF2 INR at 4.4%

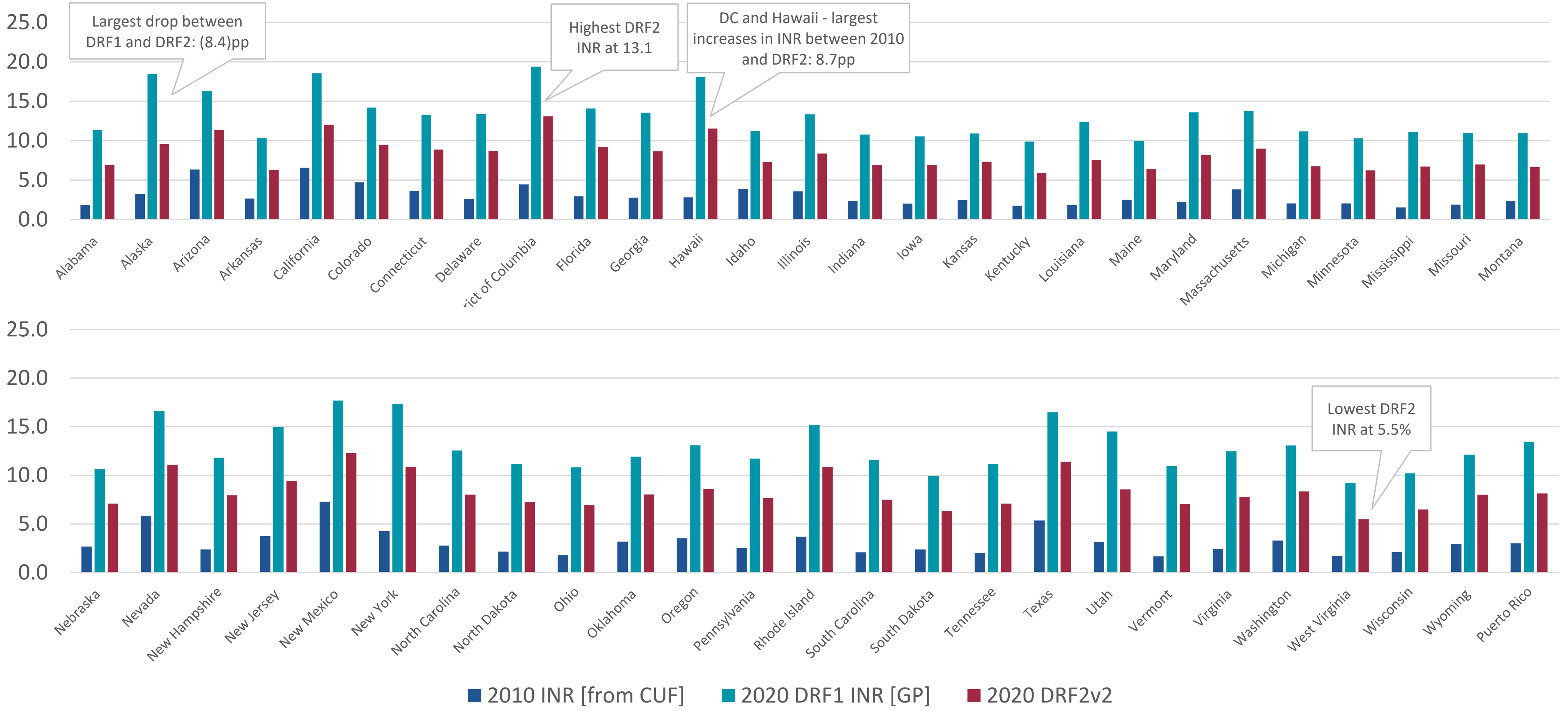
Pre-decisional - Internal Use Only - Not for Public Distribution - CBDRB-FY21-ACSO002-001

Total, Household, and Group Quarters INR for Hispanic Origin – 2010 Census and 2020 DRF2v2



Pre-decisional - Internal Use Only - Not for Public Distribution - CBDRB-FY21-ACSO002-001

Race INR - 2010 Census, 2020 DRF1 Good Persons, and 2020 DRF2 Selected Persons
Universe: Total Population

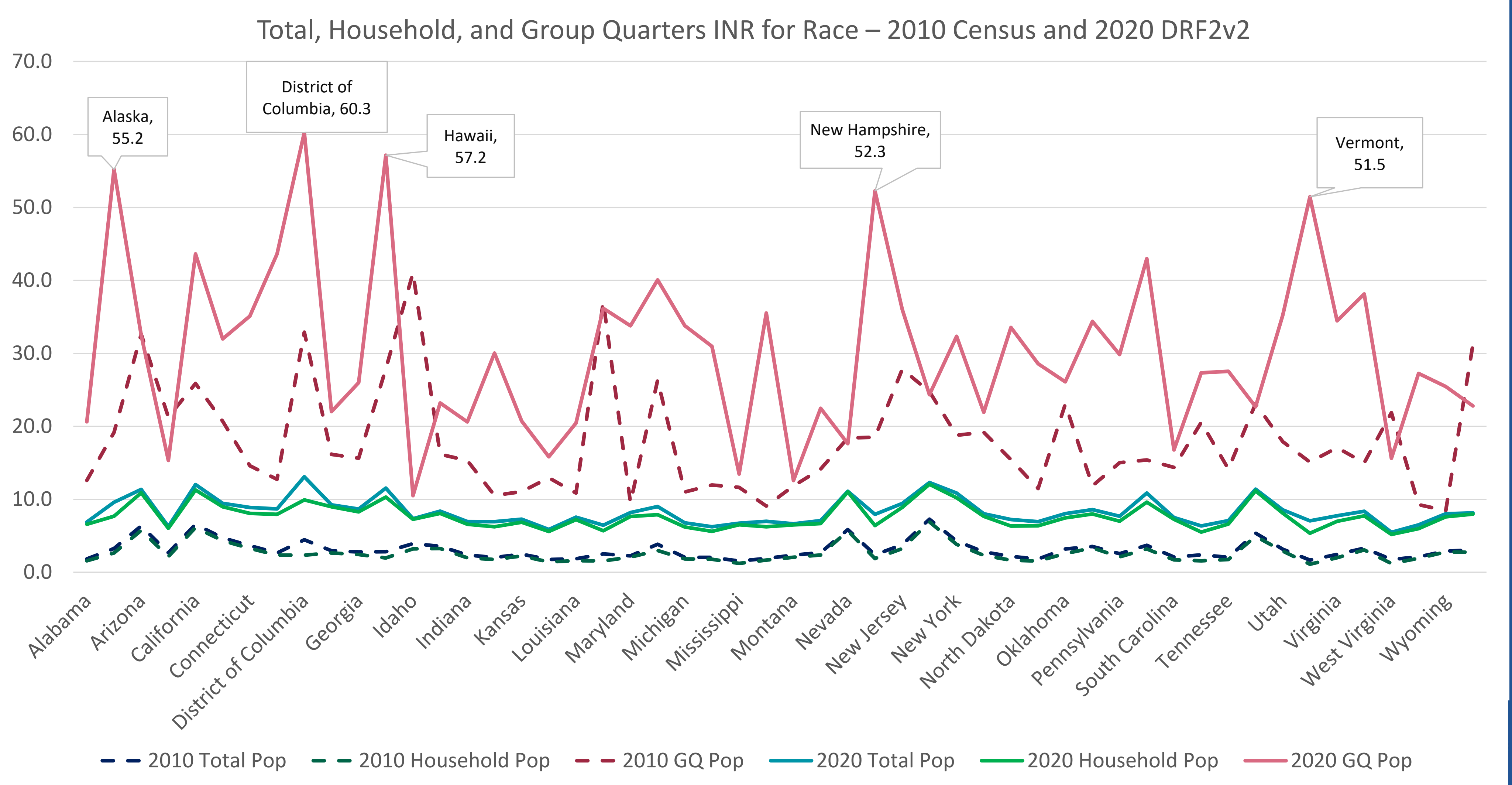


Largest drop between DRF1 and DRF2: (8.4)pp

Highest DRF2 INR at 13.1

DC and Hawaii - largest increases in INR between 2010 and DRF2: 8.7pp

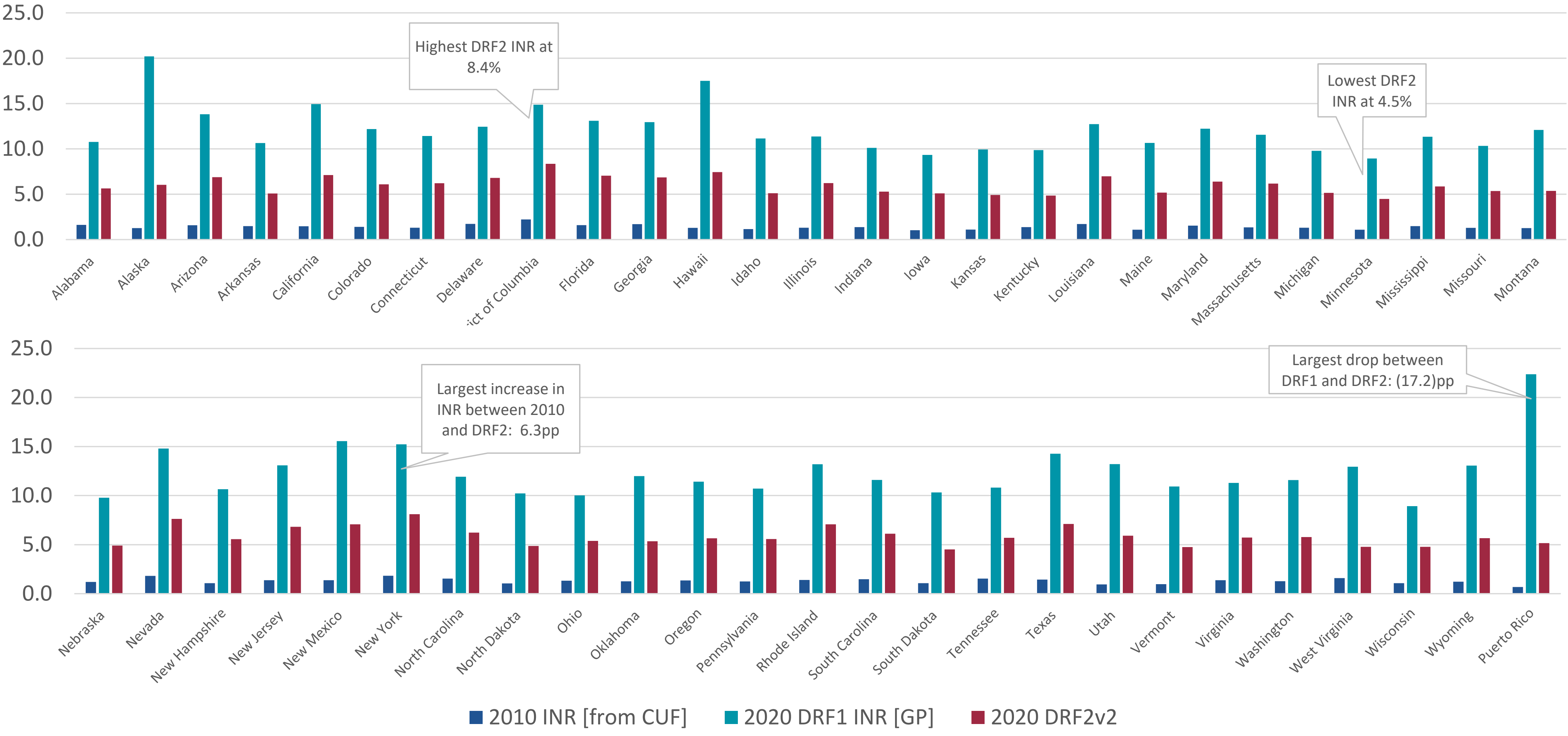
Lowest DRF2 INR at 5.5%



Total, Household, and Group Quarters INR for Race – 2010 Census and 2020 DRF2v2

Pre-decisional - Internal Use Only - Not for Public Distribution - CBDRB-FY21-ACSO002-001

Relationship INR 2010 Census, 2020 DRF1 Good Persons, and 2020 DRF2 Selected Persons
Universe: Household Population


Highest DRF2 INR at 8.4%

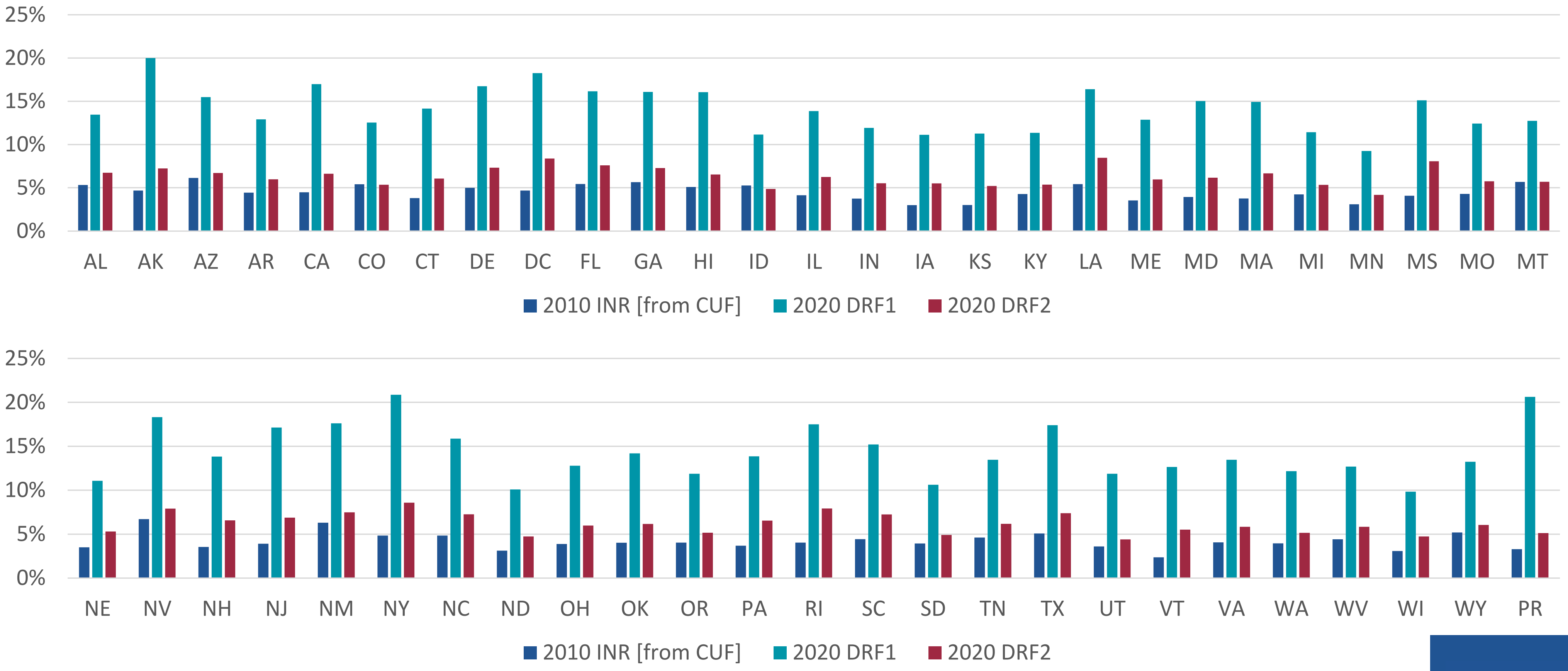
Lowest DRF2 INR at 4.5%

Largest increase in INR between 2010 and DRF2: 6.3pp

Largest drop between DRF1 and DRF2: (17.2)pp

Pre-decisional - Internal Use Only - Not for Public Distribution - CBDRB-FY21-ACSO002-001

Tenure Item Nonresponse (INR) – 2010 and 2020
Universe: Total Housing Units DRF2 v2



Caveats & Qualifications

Caveats & Qualifications:

- Tabulations come from DRF2 version 2 and do not reflect any further changes made between version 2 and the final DRF2.
- Tabulations are preliminary analyses meant for internal use. The set of observations used to produce the tabulations and the coding rules used to define item non-response may therefore differ slightly from item non-response tabulations produced by other analysts.
- Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. The U.S. Census Bureau reviewed this data product for unauthorized disclosure of confidential information and approved the disclosure avoidance practices applied to this release. CBDRB-FY21-ACSO002-001

EXHIBIT 7

June 29 Email—Plaintiff's Challenged Redactions

From: Ken Daines <KDaines@HoltzmanVogel.com>
Sent: Tuesday, June 29, 2021 8:06 PM
To: Kossak, Jonathan (CIV) <Jonathan.Kossak@usdoj.gov>
Cc: Jason Torchinsky <jtorchinsky@HoltzmanVogel.com>
Subject: RE: Fair Lines Am. Found. v. Commerce, No. 1:21-cv-01361 (DDC)

Jonathan,

As we discussed, I am attaching a pdf with 115 redacted pages pulled from the Bureau's 991-page production where it is most apparent (and in several cases indisputable) that summary statistical information was improperly redacted. Without providing an exhaustive description of our rationale for challenging each page, here are some examples where redaction under Title 13 was improper (along with corresponding page numbers from the pdf we are providing):

- **GQTYPCUR Statistical Summary Pages (pp. 1-77):** Here it is clear that statistical summary data is redacted, including the Min, Q1-3, Max, and in some cases the Mean, Range, and Std Dev. What appears to be histograms at the bottom of each page are also improperly redacted. The information from these pages are improperly redacted under 13 U.S.C. § 8(b) because it is only *derived from* raw data, but does not include the numbers that were furnished by any particular establishment or individual to the Bureau, and would not lead to disclosure of such data or include identifying information. Furthermore, the data is inconsistently redacted, suggesting that an arbitrary method was used; for instance, on page 44, every piece of data is redacted, even though the same types of data on the previous and subsequent pages are not redacted. On some pages the range and the mean are fully included, while other pages have them partially or fully redacted.

- **County Distribution of 2020 Census – GQ Person Ratios Before and After Imputation (pp. 79-82, 104-105) –** The title of these pages makes clear that group quarters distribution numbers are shown on the county level, including summary statistical information or tabulations that do not disclose any raw data reported by particular respondents.
- **Pages 83-89 –** Redacted information includes summary statistical information that is not the originally reported raw data, including Mean, Std. Dev, Minimum, Maximum, and Median, Mode, 25th and 75th Percentiles.
- **Pages 90-91** – Histograms are redacted, but no reason to believe these include raw data reported by particular respondents.
- **Group Quarters Imputation Methodology (p. 92)** – "Median Good People Count" is summary or tabulated data, not data that was originally reported or identifying data.
- **District of Columbia and South Carolina tables/charts (pp. 94-95)** – The categories of data described are clearly summary in nature, and would not lead to disclosure of any particular respondent's reported data. E.g., for D.C. it includes a row titled "2020 DRF1 Total Population" that is improperly redacted.
- **"Summarizing the Map" (p. 97)** – The numbers in this document by its own description, "summarizing," are nothing more than summary data. E.g., one redacted number pertains to the number of tracts that have a percentage decline of 90% or more, etc. But none of these include raw data as it was reported by individual respondents.
- **Census Tracts with 100% Decline from 2013-2017 ACS (p. 98)** – Here the Bureau could provide the state-, county-, and tract-level information while omitting the identifying facility names. The same is true for other pages with Census tracts data, including **pages 100-101**.
- **Pages 106-108** – These also appear to be summary statistics based on the table format, although it is admittedly difficult to tell based on the full redaction.
- **Tracts with Largest Number of Nursing Home People Found in a GQ (pp. 109-114)** – The state-, county-, and tract-level data is summary statistical information that does not disclose information reported by any particular respondent.
- **10 Counties with Highest % Enrolled (p. 115)** – The Bureau can provide the percentage, county- and state-level information, without providing particular university information.

Please note that by providing these examples, including the pdf, we are not waiving our right to challenge improper redactions on the other redacted pages, many of which are fully redacted which makes it impossible to tell whether redaction was improper.

Also, as discussed on the call, we look forward to your update this week regarding the post-December 2020 documents and the 2600 emails (including the number of pages).

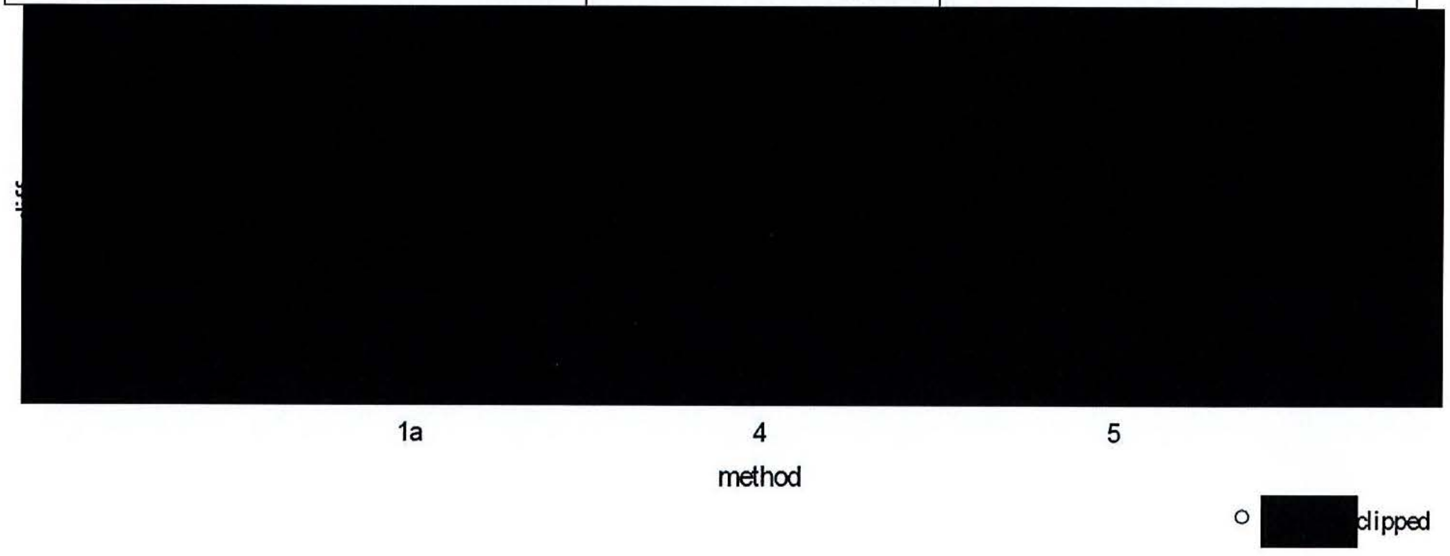
Thank you,

Ken

Ken Daines
KDaines@HoltzmanVogel.com // www.HoltzmanVogel.com

GQTYPCUR=101

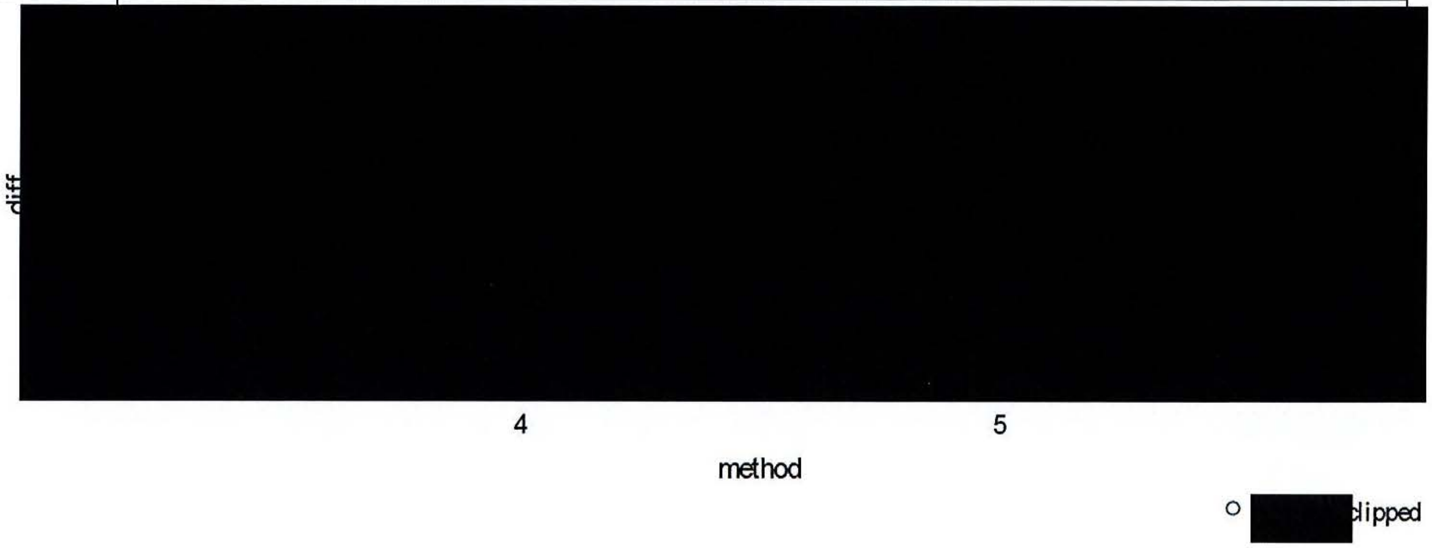
Statistics by GQtype			
N	50	900	0
Min	[redacted]	[redacted]	.
Nlow	<15	0	0
Q1	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	.
Mean	-18	74.[redacted]	.
Q3	[redacted]	[redacted]	.
Nhigh	<15	40	0
Max	[redacted]	[redacted]	.
Nout	<15	40	0
Range	850	3900	.
Std Dev	110.[redacted]	248[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=102

Statistics by GQtype		
N	250	0
Min	[redacted]	.
Nlow	<15	0
Q1	[redacted]	.
Q2	[redacted]	.
Mean	645[redacted]	.
Q3	[redacted]	.
Nhigh	<15	0
Max	[redacted]	.
Nout	20	0
Range	2800	.
Std Dev	537[redacted]	.

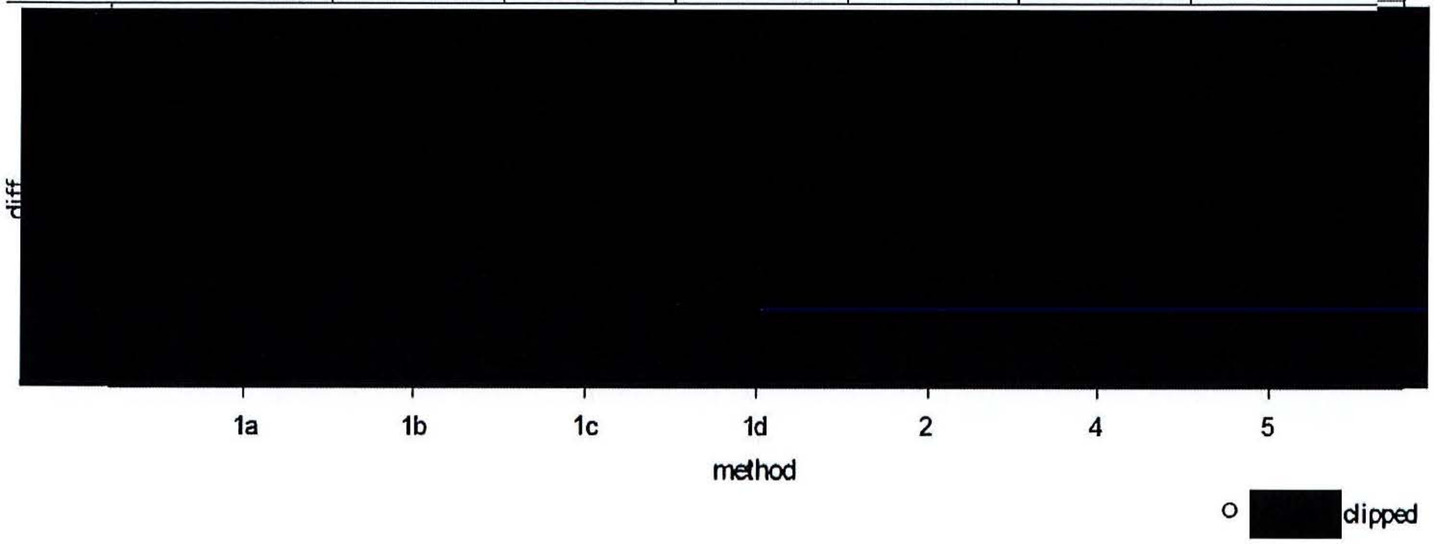

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=103

Statistics by GQtype

N	4700	550	550	550	550	8000	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	250	30	30	30	30	400	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0[redacted]	-7.[redacted]	-0[redacted]	-7.[redacted]	-0.[redacted]	-83[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	250	30	30	30	30	400	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	450	50	50	50	50	800	0
Range	3300	1500	2700	1500	1900	4700	.
Std Dev	174.[redacted]	171.[redacted]	170.[redacted]	172.[redacted]	138.[redacted]	199.[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=104

Statistics by GQtype

N	2700	2000	2000	2000	2000	2800	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	150	100	100	100	100	150	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	1.[redacted]	-1.[redacted]	-1.[redacted]	-3.[redacted]	-0.[redacted]	-0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	150	100	100	100	100	150	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	250	200	200	200	200	250	0
Range	3000	3600	3100	3600	6000	4100	.
Std Dev	118.[redacted]	136.[redacted]	114.[redacted]	137.[redacted]	145.[redacted]	271.[redacted]	.

diff

[redacted]

1a 1b 1c 1d 2 4 5

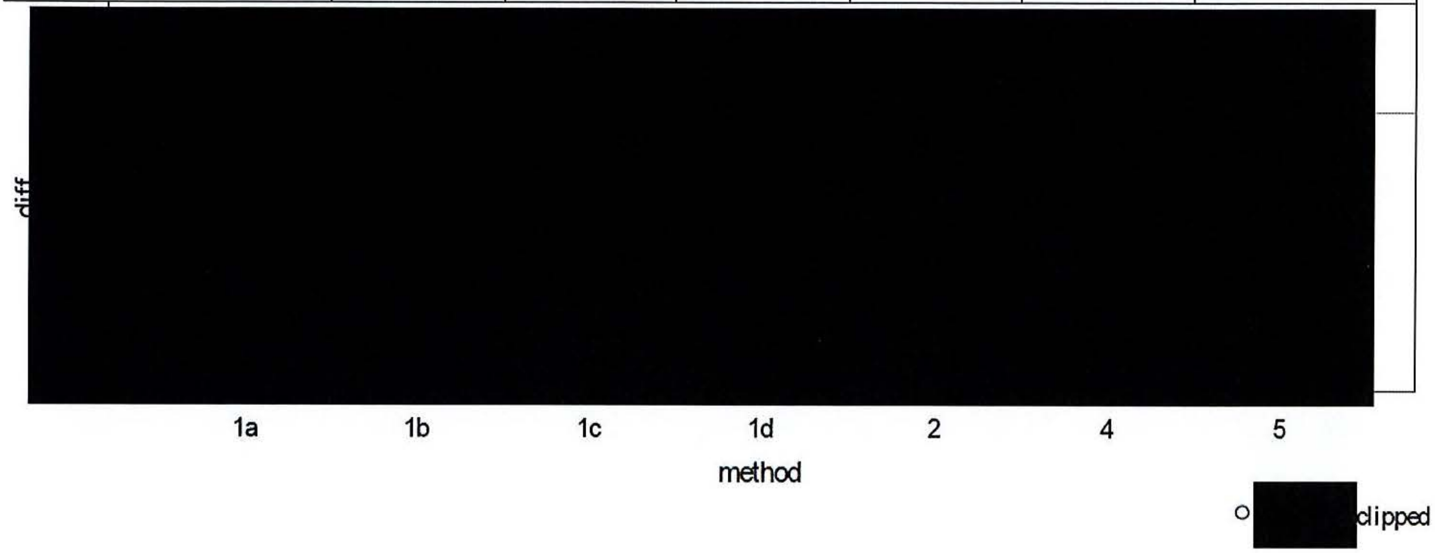
method

[redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

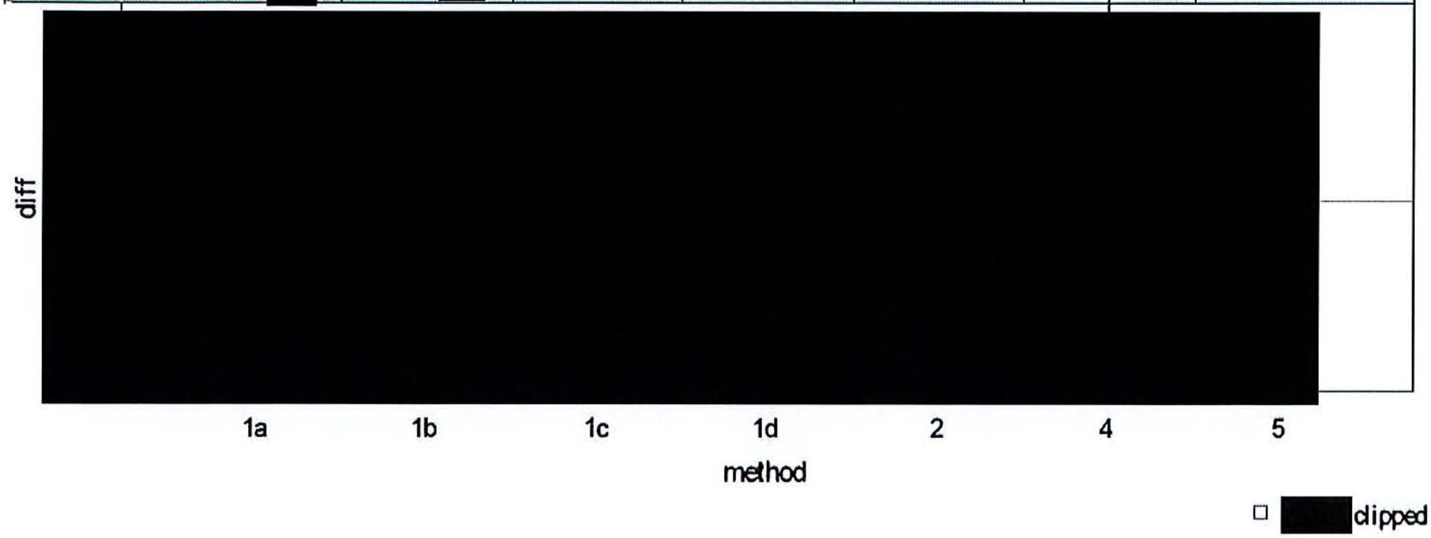
GQTYPCUR=105

Statistics by GQtype							
N	700	350	350	350	350	850	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	30	20	20	20	20	40	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-2.[redacted]	2.[redacted]	-3.[redacted]	-0.[redacted]	0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	30	20	20	20	20	40	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	70	30	30	30	30	80	0
Range	2900	750	600	750	450	3000	.
Std Dev	108.[redacted]	61.[redacted]	63.[redacted]	65.[redacted]	46.[redacted]	164[redacted]	.


[redacted]

GQTYPCUR=106

Statistics by GQtype							
N	20	<15	<15	<15	<15	20	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	0	0	0	0	0	0	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-22.[redacted]	-1.[redacted]	24	16[redacted]	-1.[redacted]	-18.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	0	0	0	0	0	0	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	0	0	0	0	0	0	0
Range	250	150	150	100	200	450	.
Std Dev	63.[redacted]	57.[redacted]	55.[redacted]	45[redacted]	73.[redacted]	36.[redacted]	.

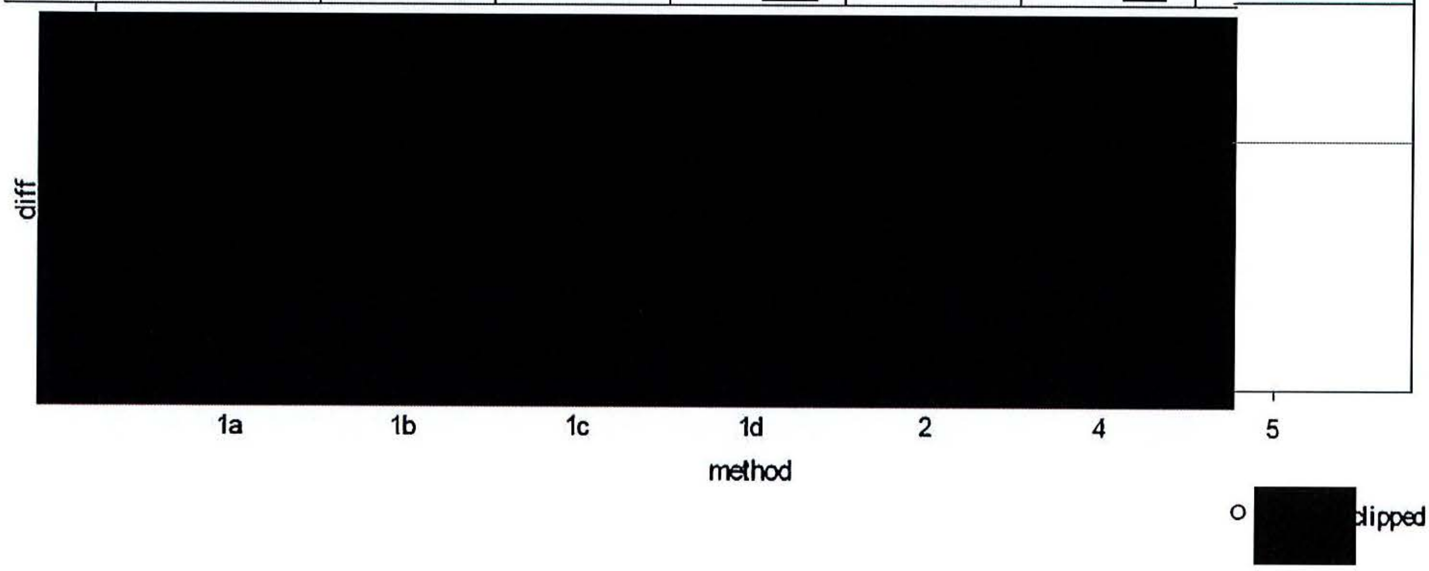

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=201

Statistics by GQtype

N	3000	1400	1400	1400	1400	3100	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	150	60	70	60	70	100	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-0.[redacted]	-0[redacted]	-0[redacted]	-0.[redacted]	0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	150	70	70	60	70	150	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	300	150	150	100	150	250	0
Range	250	150	150	150	150	200	.
Std Dev	8.[redacted]	9.[redacted]	8.[redacted]	9.[redacted]	8.[redacted]	12.[redacted]	.


[redacted]

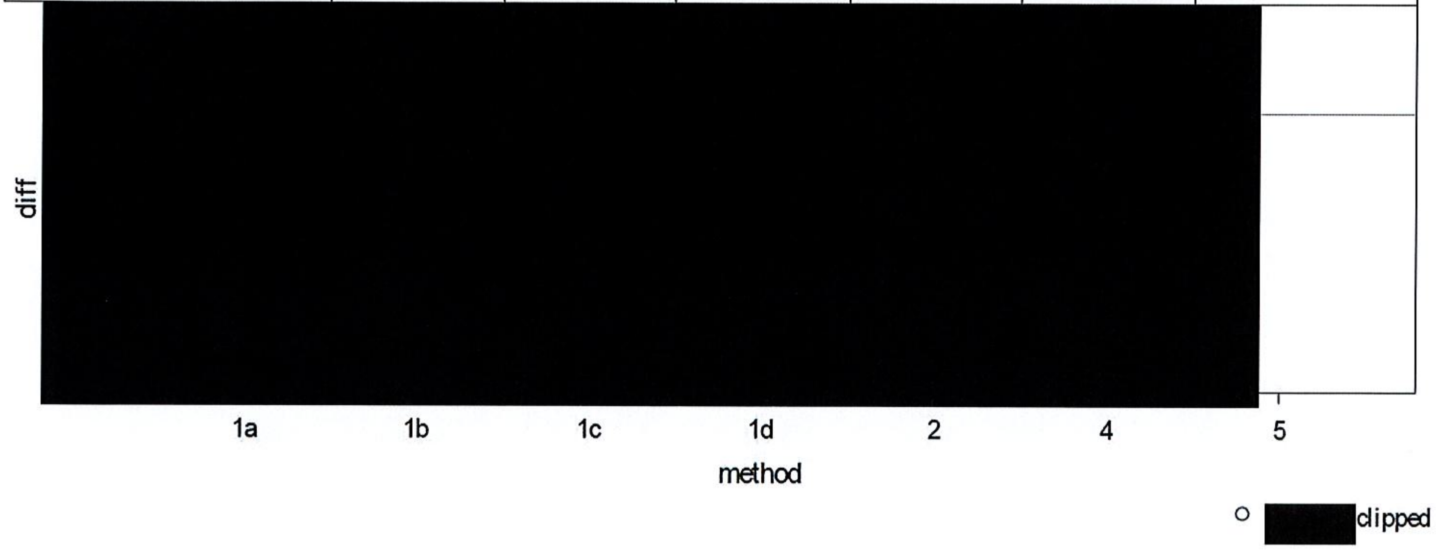
[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=202

Statistics by GQtype

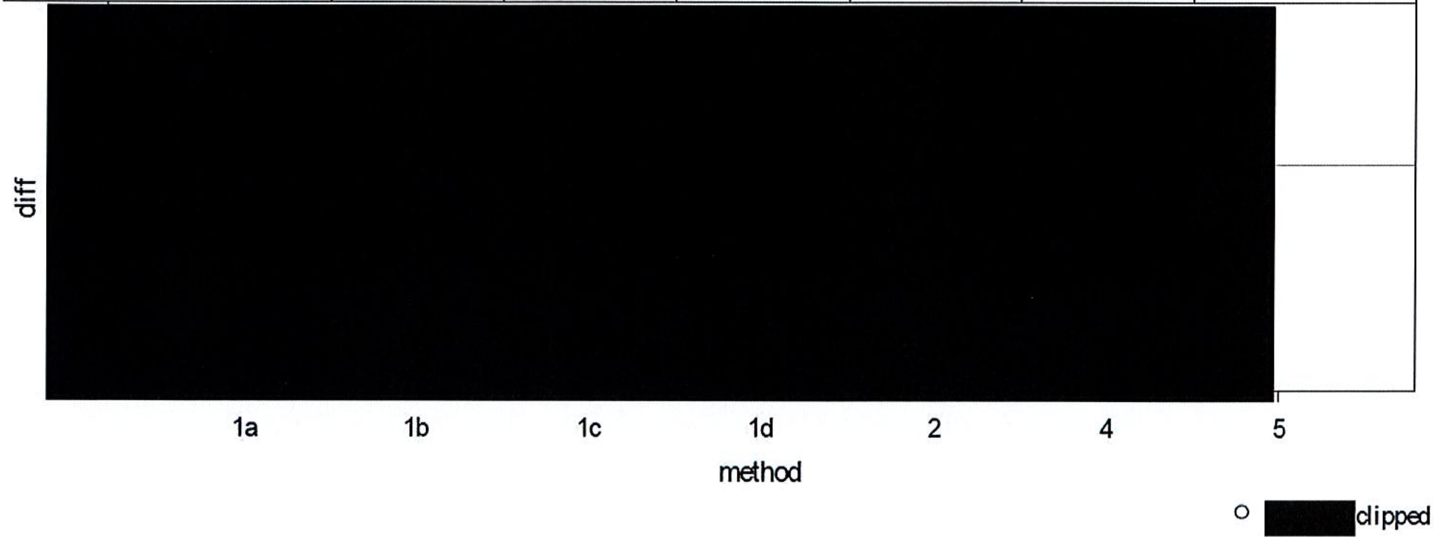
N	1600	850	850	850	850	1700	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	80	40	40	40	40	80	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0[redacted]	-2.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	70	40	40	40	40	80	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	150	80	80	80	80	150	0
Range	150	150	200	150	150	300	.
Std Dev	11.[redacted]	13.[redacted]	13.[redacted]	13.[redacted]	11.[redacted]	22.[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=203

Statistics by GQtype							
N	1000	550	550	550	550	1100	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow							0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-3.[redacted]	-2[redacted]	-3.[redacted]	-0.[redacted]	0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	50	30	30	30	30	50	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	100	60	60	50	60	100	0
Range	700	650	650	600	250	350	.
Std Dev	26.[redacted]	32.[redacted]	28.[redacted]	30.[redacted]	19[redacted]	29[redacted]	.

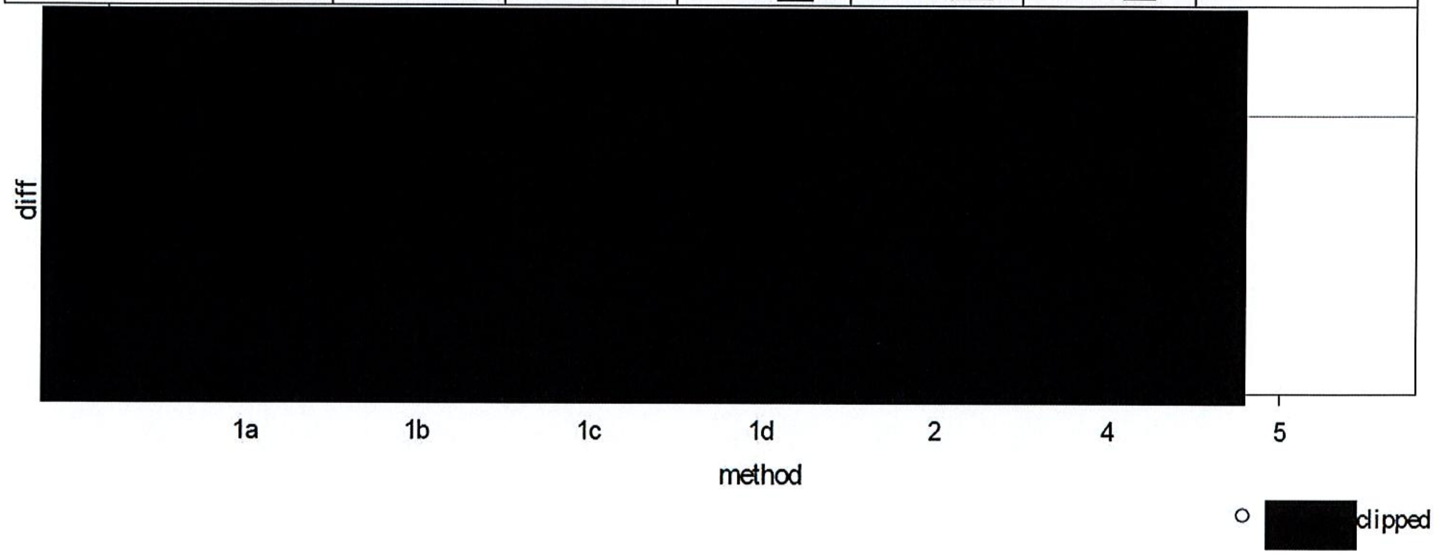


DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=301

Statistics by GQtype

N	24000	14500	14500	14500	14500	24500	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	1200	700	700	700	700	1200	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-0.[redacted]	-0[redacted]	0.[redacted]	-0.[redacted]	-10.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	1200	700	700	700	700	1200	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	2400	1400	1400	1400	1400	2400	0
Range	1400	1500	1100	1500	1100	1400	.
Std Dev	26.[redacted]	27[redacted]	28.[redacted]	27.[redacted]	25.[redacted]	52.[redacted]	.

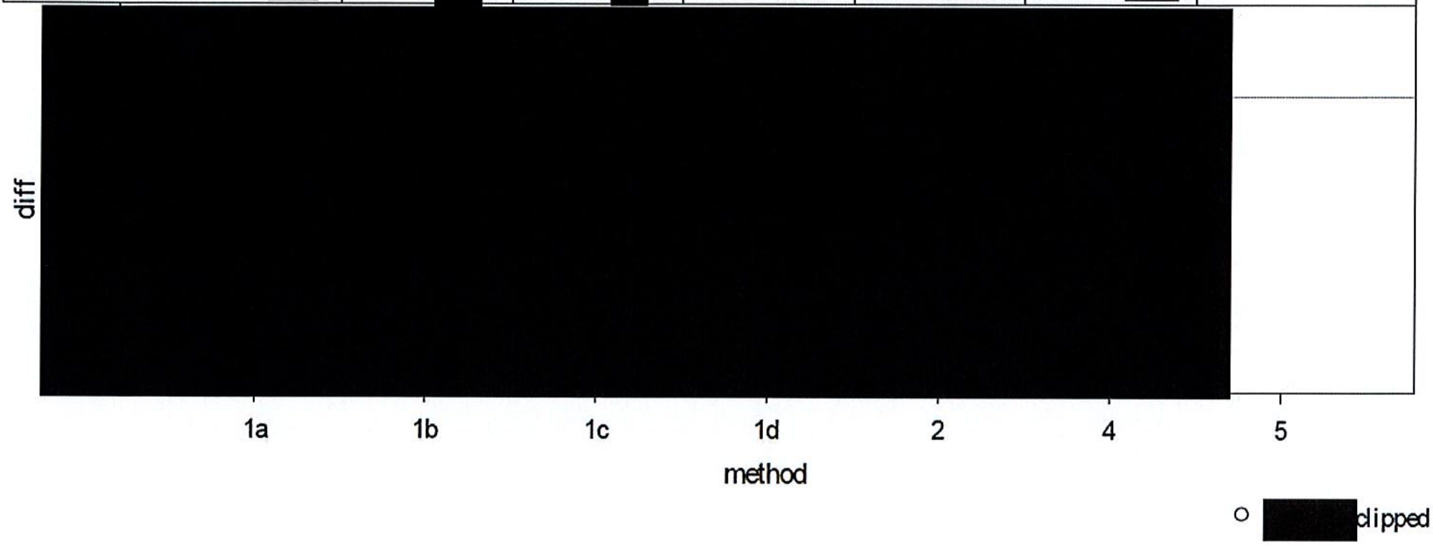

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=401

Statistics by GQtype

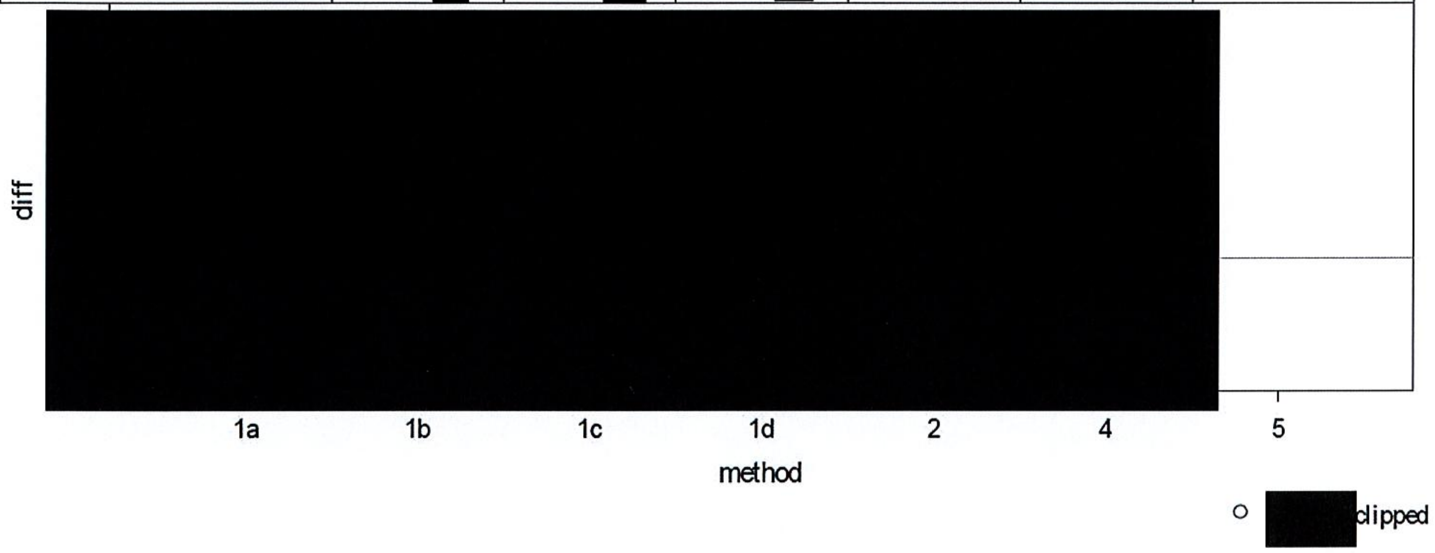
N	650	250	250	250	250	700	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	30	<15	<15	<15	<15	30	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-5.[redacted]	-3.[redacted]	-0[redacted]	-0.[redacted]	-0.[redacted]	-25.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	30	<15	<15	<15	<15	30	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	60	20	20	20	20	70	0
Range	1400	1400	800	1400	600	1100	.
Std Dev	65.[redacted]	82[redacted]	62.[redacted]	78.[redacted]	46.[redacted]	89.[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=402

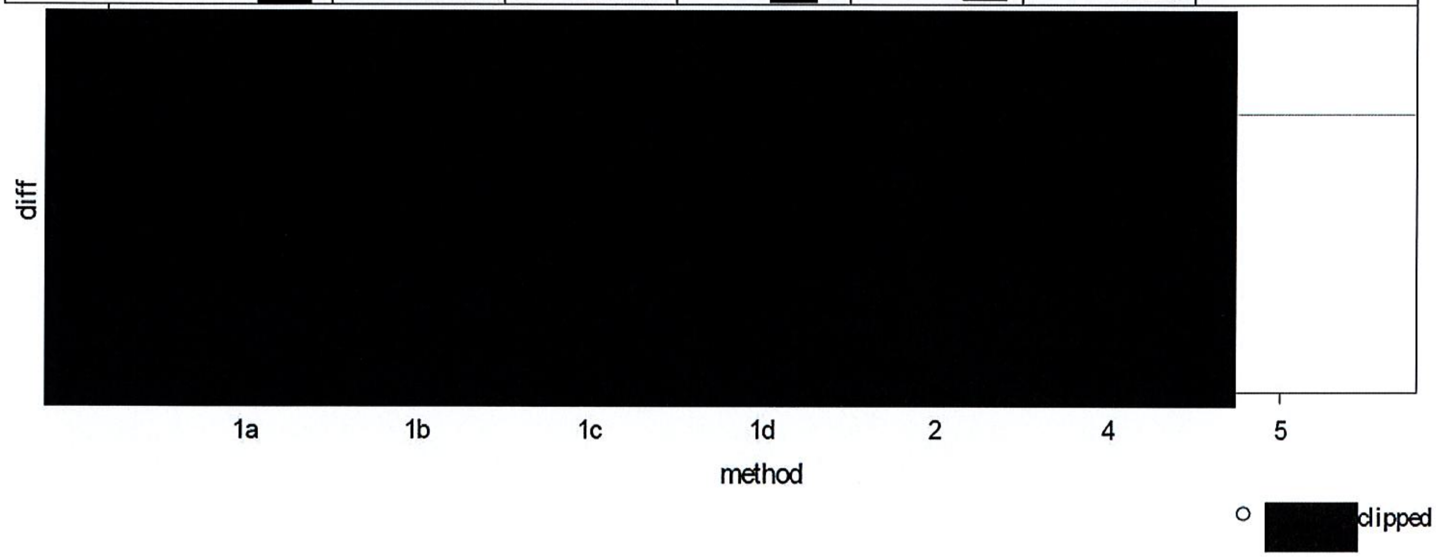
Statistics by GQtype							
N	200	70	70	70	70	200	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-6.[redacted]	-2.[redacted]	-6[redacted]	-4.[redacted]	-0.[redacted]	-6.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	20	<15	<15	<15	<15	20	0
Range	350	200	200	400	300	600	.
Std Dev	38.[redacted]	26.[redacted]	33.[redacted]	42.[redacted]	30.[redacted]	59.[redacted]	.



DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=403

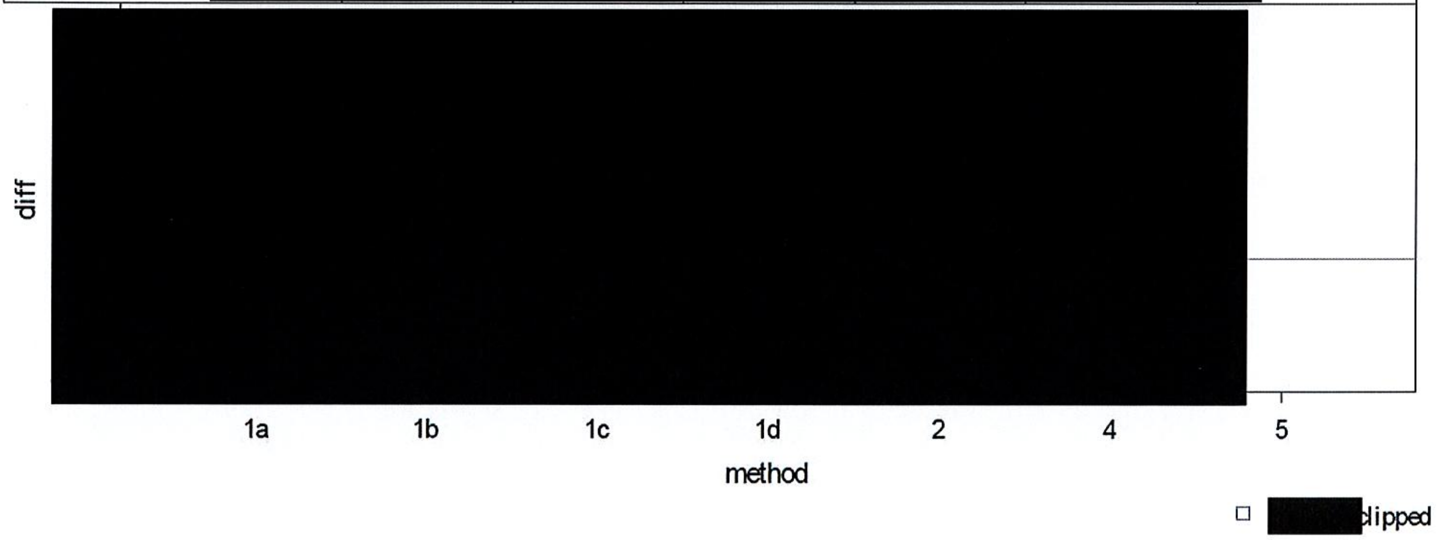
Statistics by GQtype							
N	450	200	200	200	200	450	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	20	<15	<15	<15	<15	20	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0.[redacted]	-0.[redacted]	-0.[redacted]	0.[redacted]	-0.[redacted]	1[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	20	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	40	20	20	20	20	40	0
Range	150	90	100	90	150	400	.
Std Dev	12.[redacted]	9.[redacted]	11.[redacted]	9.[redacted]	10.[redacted]	39.[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=404

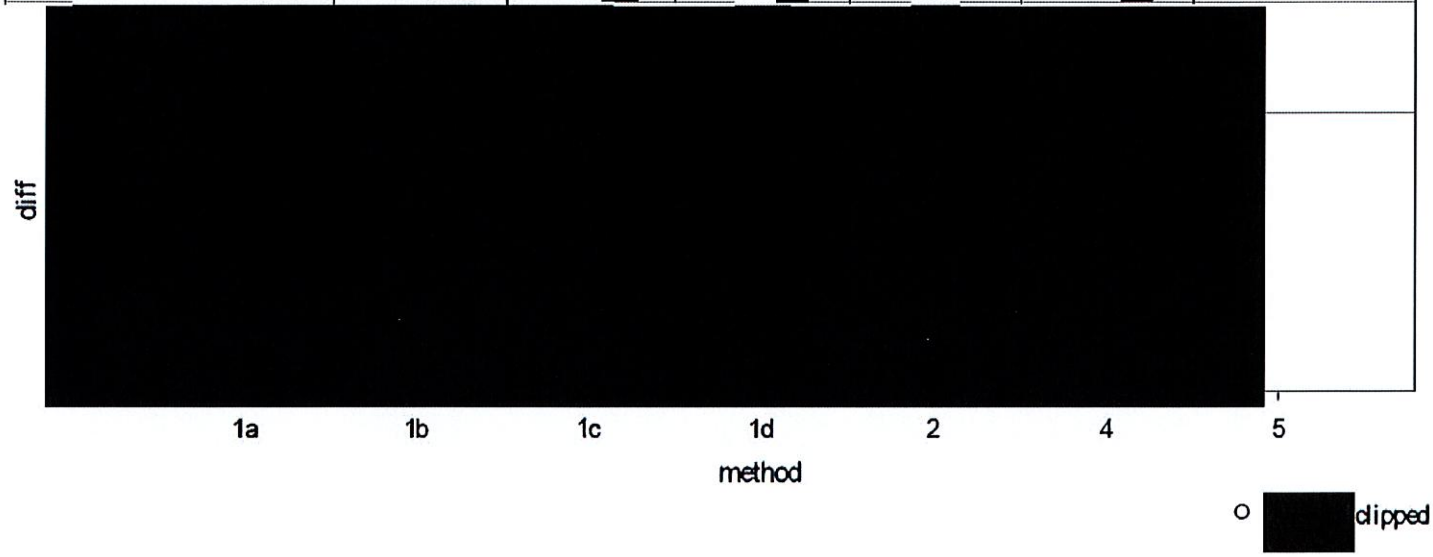
Statistics by GQtype							
N	<15	<15	<15	<15	<15	<15	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.



DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=405

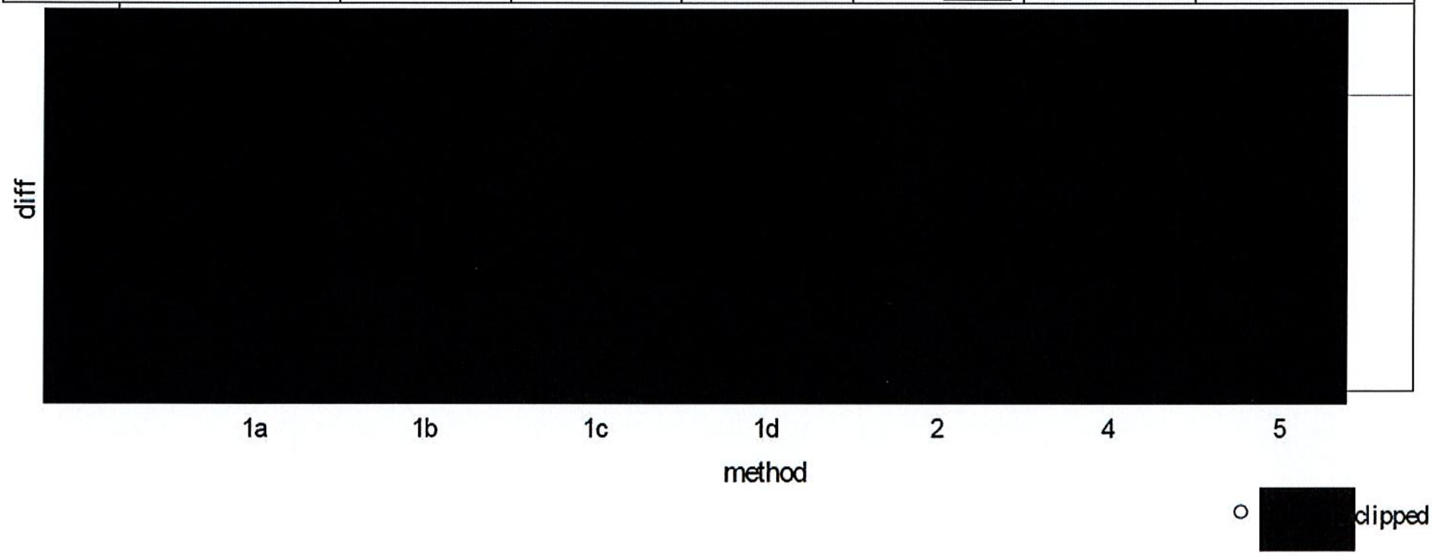
Statistics by GQtype							
N	450	250	250	250	250	500	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	20	<15	<15	<15	<15	20	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1.[redacted]	-3.[redacted]	-0.[redacted]	-3.[redacted]	-0.[redacted]	-6.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	20	<15	<15	<15	<15	20	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	40	20	20	20	20	50	0
Range	350	300	300	300	200	550	.
Std Dev	21.[redacted]	30.[redacted]	25.[redacted]	30.[redacted]	21.[redacted]	39.[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=501

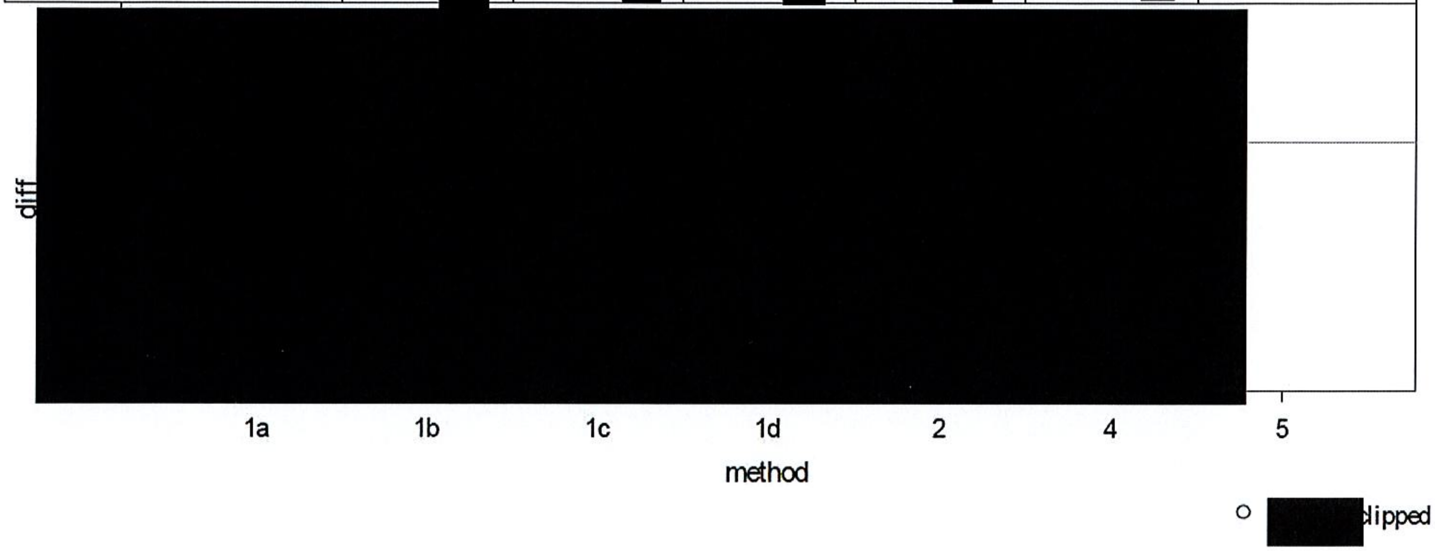
Statistics by GQtype							
N	27000	13000	13000	13000	13000	28000	28000
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nlow	1300	650	650	650	650	1300	1400
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	0[redacted]	-1.[redacted]	-0.[redacted]	-1.[redacted]	-0.[redacted]	-10.[redacted]	0.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nhigh	1300	650	650	650	650	1400	1400
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nout	2700	1300	1300	1300	1300	2800	2800
Range	2600	1500	1400	1500	1400	8700	6000
Std Dev	45[redacted]	42.[redacted]	38.[redacted]	42[redacted]	34[redacted]	146.[redacted]	73[redacted]


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=601

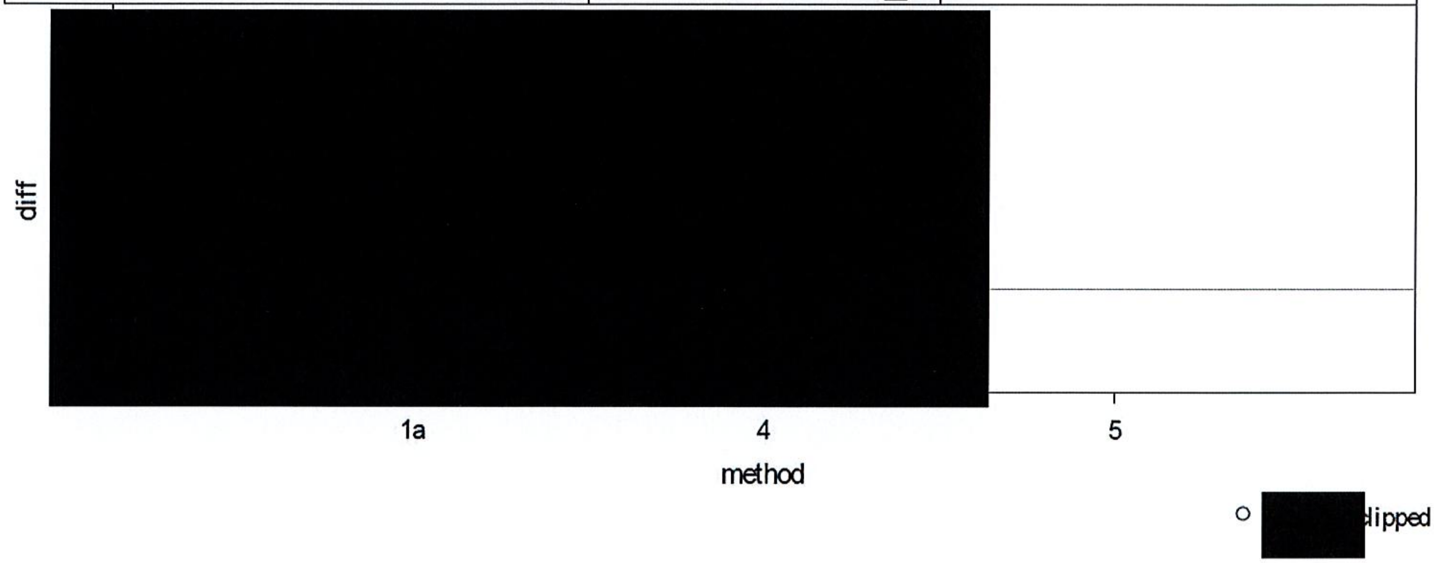
Statistics by GQtype							
N	2300	800	800	800	800	2700	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	100	40	40	40	40	100	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-4.[redacted]	-4[redacted]	-4.[redacted]	-1.[redacted]	-0.[redacted]	-33.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	100	40	40	40	40	150	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	250	80	80	80	80	250	0
Range	1900	1400	1300	1400	1200	1700	.
Std Dev	101.[redacted]	95.[redacted]	101.[redacted]	93.[redacted]	82.[redacted]	126.[redacted]	.



DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=602

Statistics by GQtype			
N	250	250	0
Min	[redacted]	[redacted]	.
Nlow	<15	<15	0
Q1	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	.
Mean	9	100.[redacted]	.
Q3	[redacted]	[redacted]	.
Nhigh	<15	<15	0
Max	[redacted]	[redacted]	.
Nout	20	20	0
Range	1600	1700	.
Std Dev	135[redacted]	262.[redacted]	.

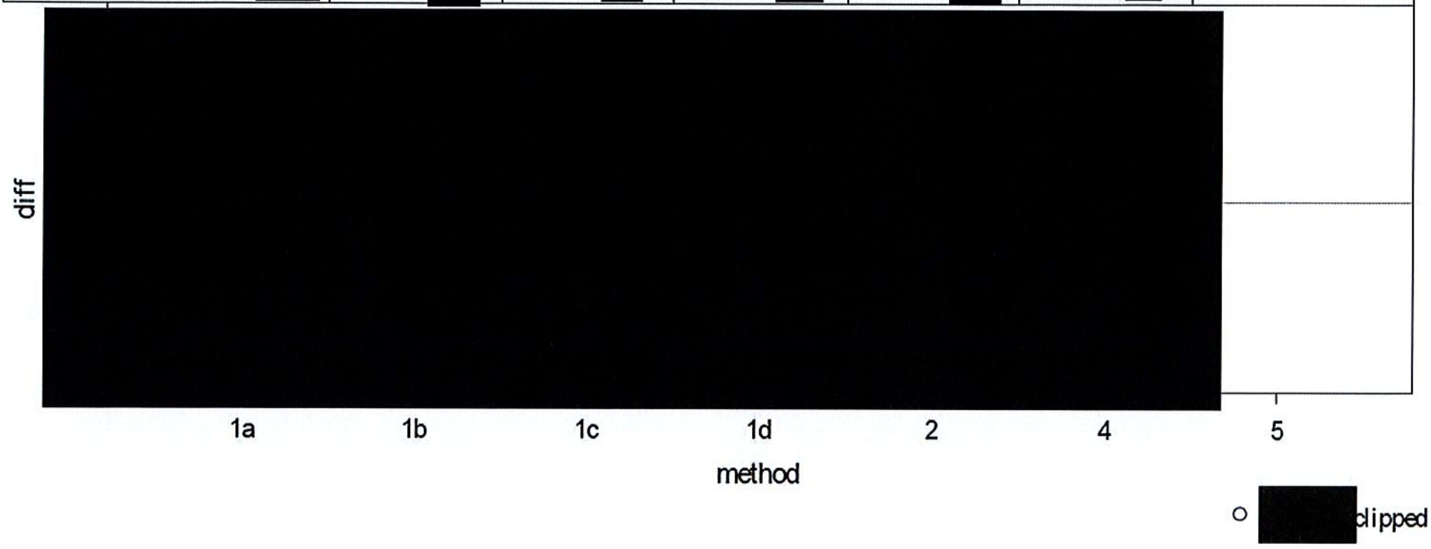

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=701

Statistics by GQtype

N	4800	1600	1600	1600	1600	5100	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	250	80	80	80	80	250	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0.[redacted]	-1.[redacted]	-0.[redacted]	-2.[redacted]	-0.[redacted]	5.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	250	80	80	80	80	250	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	450	150	150	150	150	500	0
Range	1300	800	800	800	550	1800	.
Std Dev	48.[redacted]	35.[redacted]	43.[redacted]	36.[redacted]	32.[redacted]	63.[redacted]	.

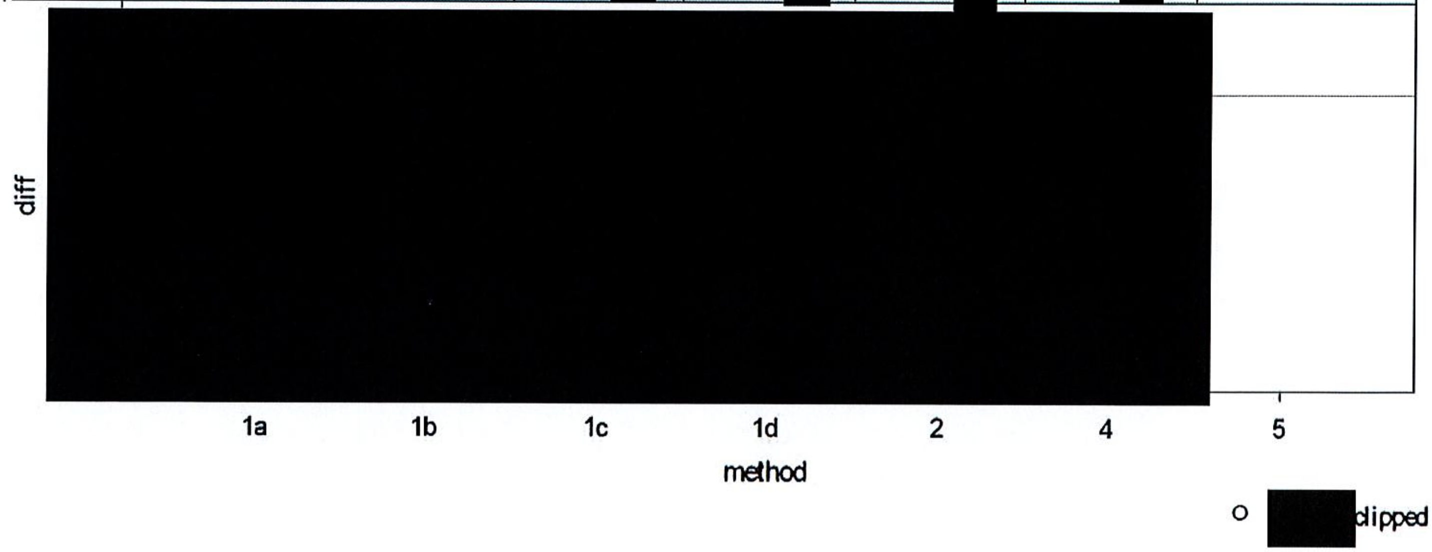


DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=702

Statistics by GQtype

N	1700	700	700	700	700	1800	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	80	40	40	40	40	80	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	4.[redacted]	-7.[redacted]	-2.[redacted]	-12.[redacted]	-0.[redacted]	0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	80	40	30	40	40	90	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	150	70	70	70	70	150	0
Range	1300	650	1700	650	850	1600	.
Std Dev	69.[redacted]	59.[redacted]	82.[redacted]	62.[redacted]	62.[redacted]	89[redacted]	.

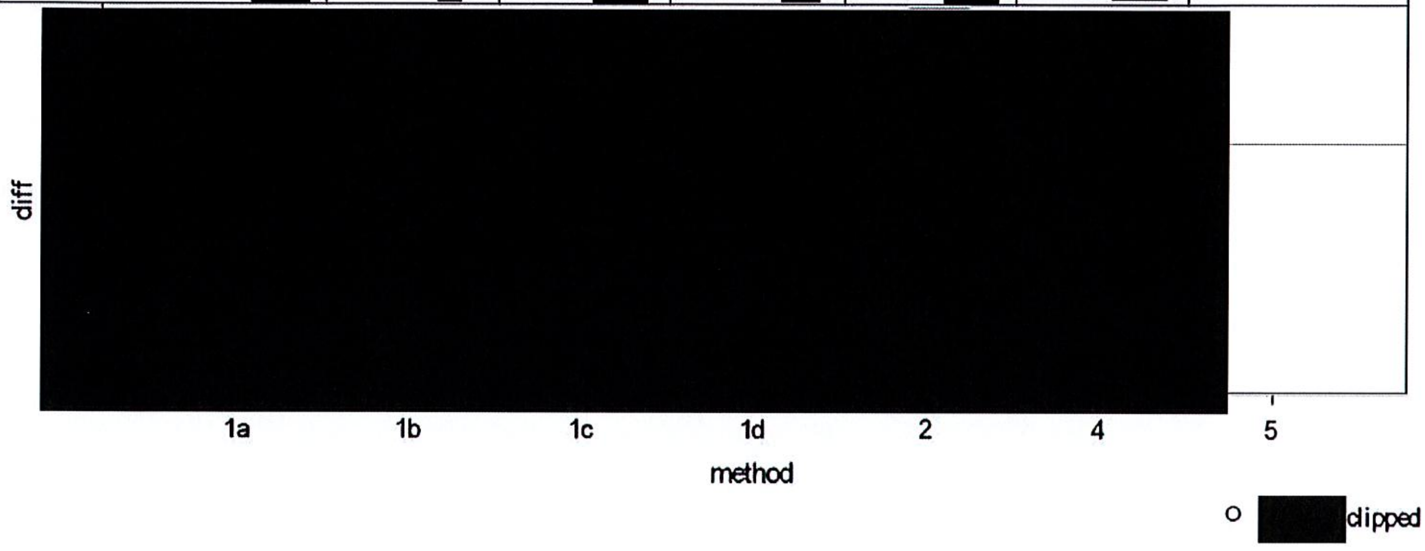

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=704

Statistics by GQtype

N	250	30	30	30	30	250	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<1	<15	<1	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	2.[redacted]	-19.[redacted]	1.[redacted]	-21.[redacted]	-0.[redacted]	13	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	30	<15	<15	<15	<15	20	0
Range	700	750	100	750	100	900	.
Std Dev	46.[redacted]	116.[redacted]	16.[redacted]	115.[redacted]	17.[redacted]	66.[redacted]	.

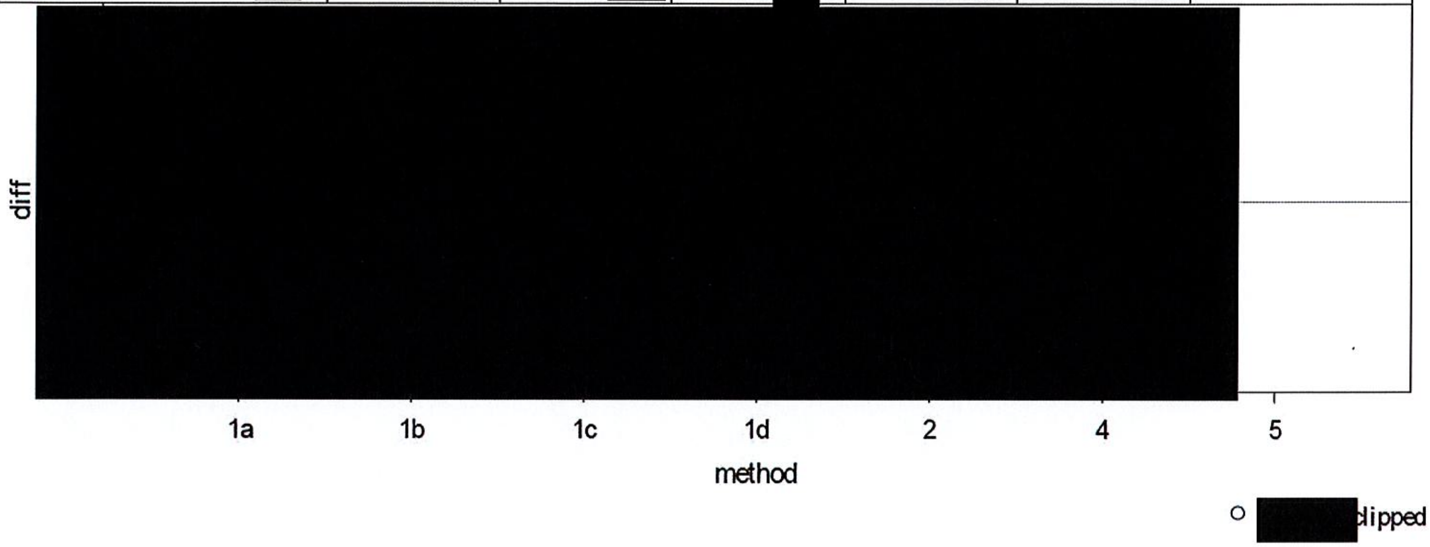

[redacted]

GQTYPCUR=706

Statistics by GQtype

N	7300	200	200	200	200	17000	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	350	<15	<15	<1	<15	800	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1.[redacted]	-4.[redacted]	-3[redacted]	-1.[redacted]	-1.[redacted]	-0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	350	<15	<15	<15	<15	800	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	700	20	20	20	20	1600	0
Range	850	200	400	150	300	600	.
Std Dev	22.[redacted]	21.[redacted]	30[redacted]	16.[redacted]	22.[redacted]	19.[redacted]	.

[redacted]

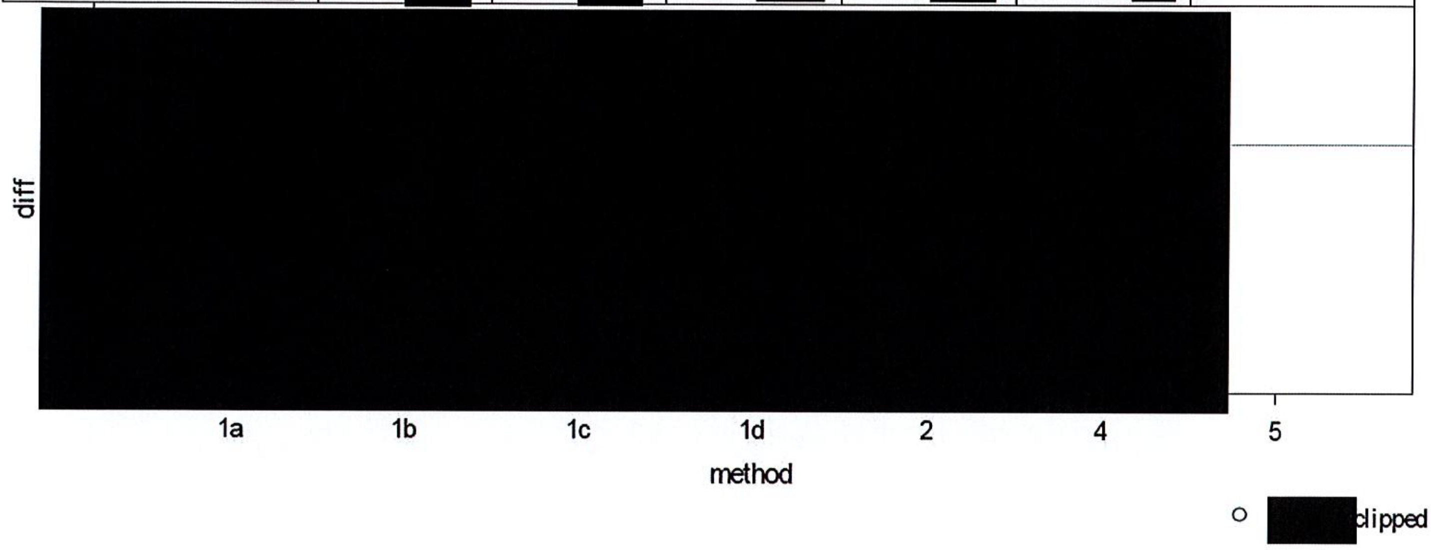


DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=801

Statistics by GQtype

N	50500	21000	21000	21000	21000	52500	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	2400	1000	1000	1000	900	1700	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-0[redacted]	-0.[redacted]	-0.[redacted]	-0[redacted]	1.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	2000	800	900	800	800	2600	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	4400	1800	1800	1800	1800	4300	0
Range	700	400	200	400	300	700	.
Std Dev	9.[redacted]	6.[redacted]	5.[redacted]	6.[redacted]	5.[redacted]	18.[redacted]	.

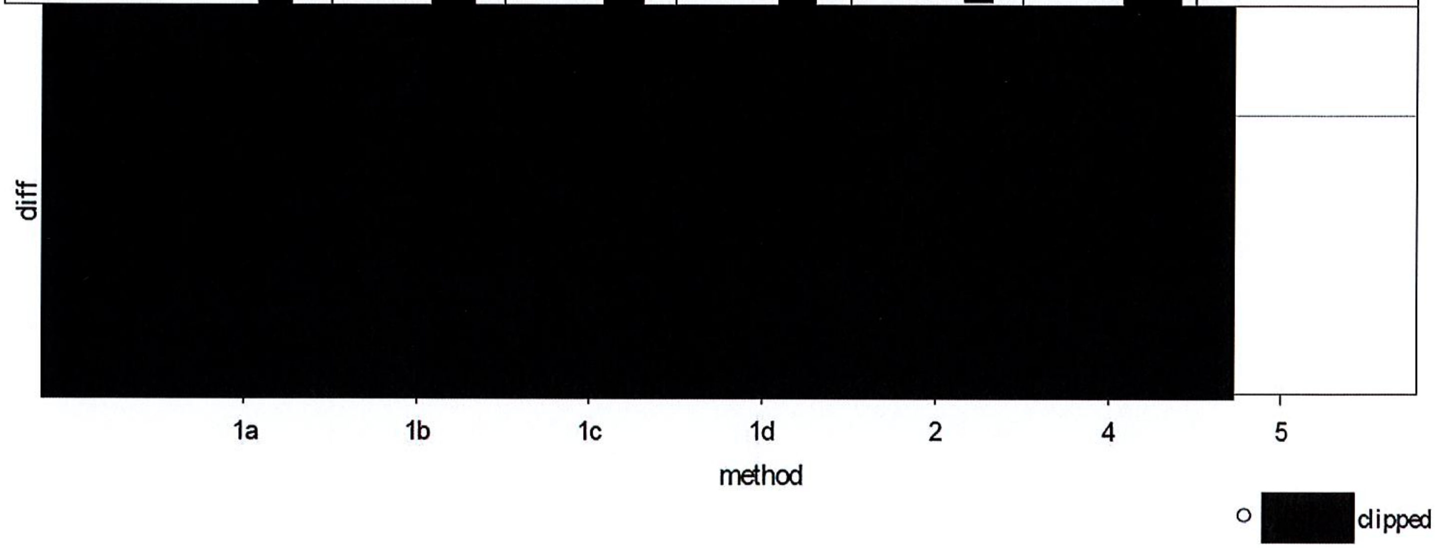

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=802

Statistics by GQtype

N	8000	3500	3500	3500	3500	8300	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	400	150	150	200	150	350	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	0.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	400	150	150	150	150	400	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	750	350	350	350	350	800	0
Range	1200	300	300	300	300	1200	.
Std Dev	17.[redacted]	13.[redacted]	13.[redacted]	13.[redacted]	12.[redacted]	31.[redacted]	.

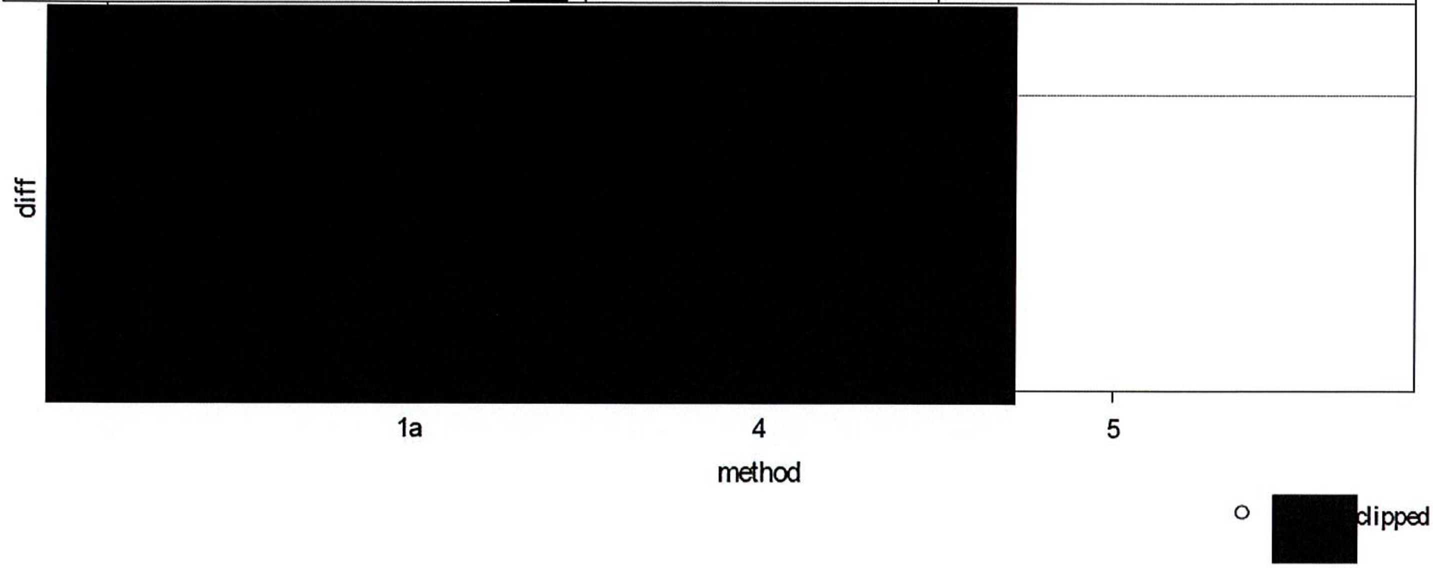

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=900

Statistics by GQtype

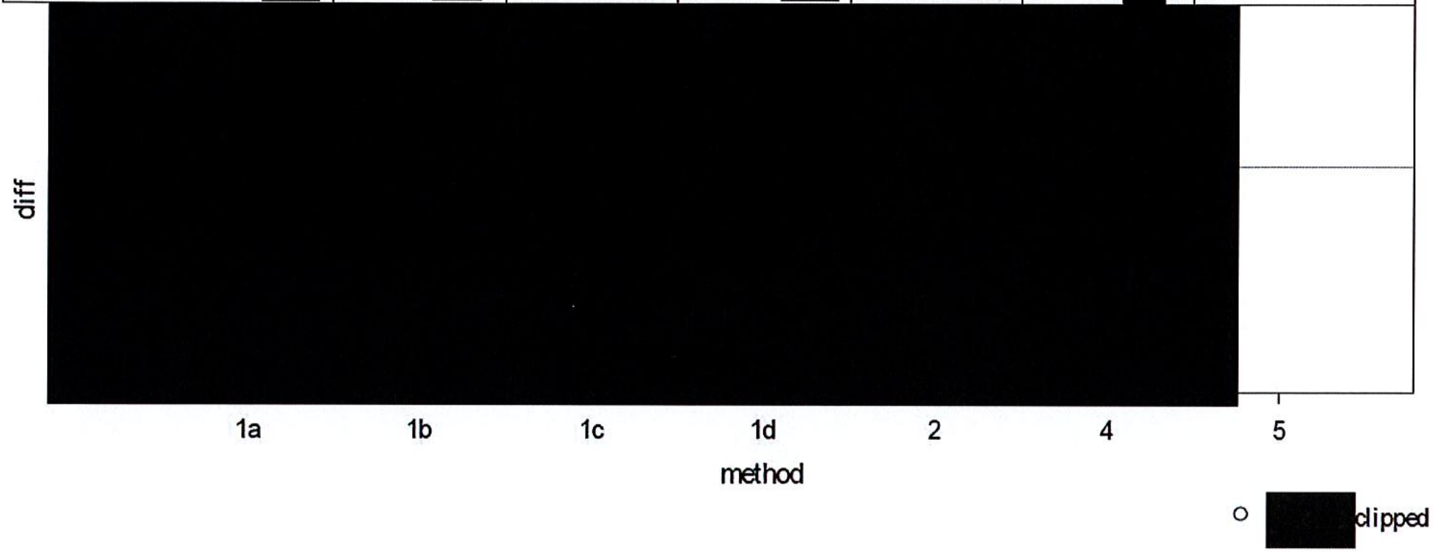
N	350	350	0
Min	[redacted]	[redacted]	.
Nlow	20	0	0
Q1	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	.
Mean	-0.[redacted]	-33.[redacted]	.
Q3	[redacted]	[redacted]	.
Nhigh	20	20	0
Max	[redacted]	[redacted]	.
Nout	30	20	0
Range	150	100	.
Std Dev	10[redacted]	12[redacted]	.



DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=901

Statistics by GQtype							
N	5300	2200	2200	2200	2200	5500	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	250	100	100	100	100	250	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	1.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	250	100	100	100	100	250	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	500	200	200	200	200	500	0
Range	1000	300	450	300	300	1100	.
Std Dev	18.[redacted]	15.[redacted]	18.[redacted]	15.[redacted]	13.[redacted]	32.[redacted]	.

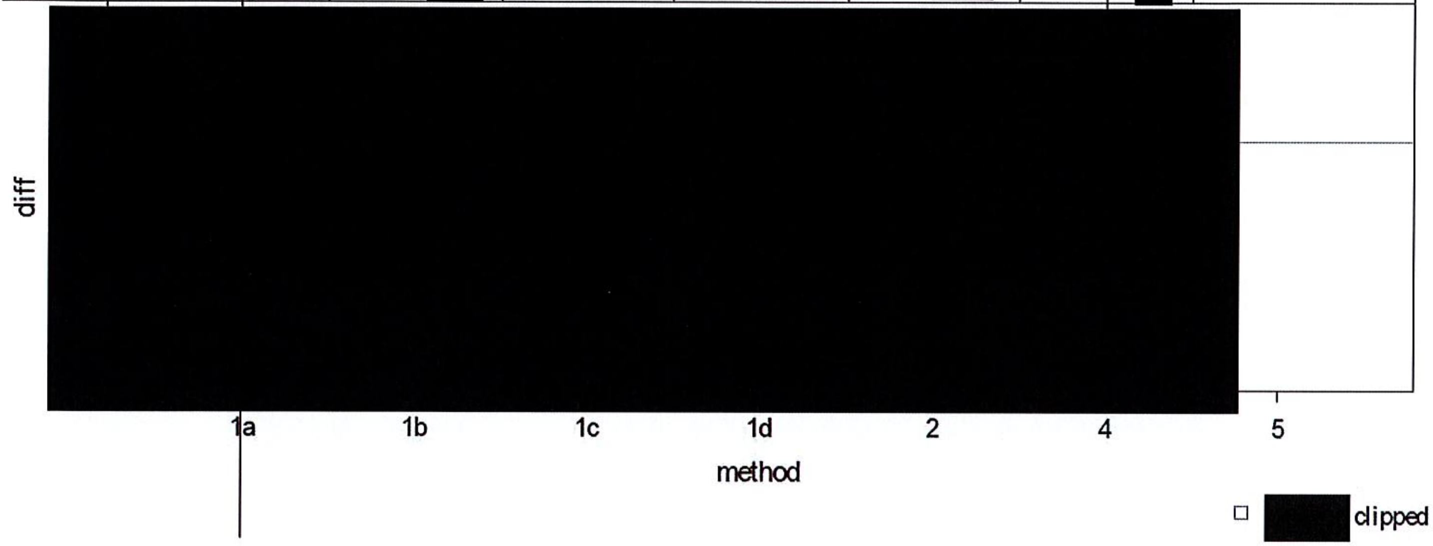

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=903

Statistics by GQtype

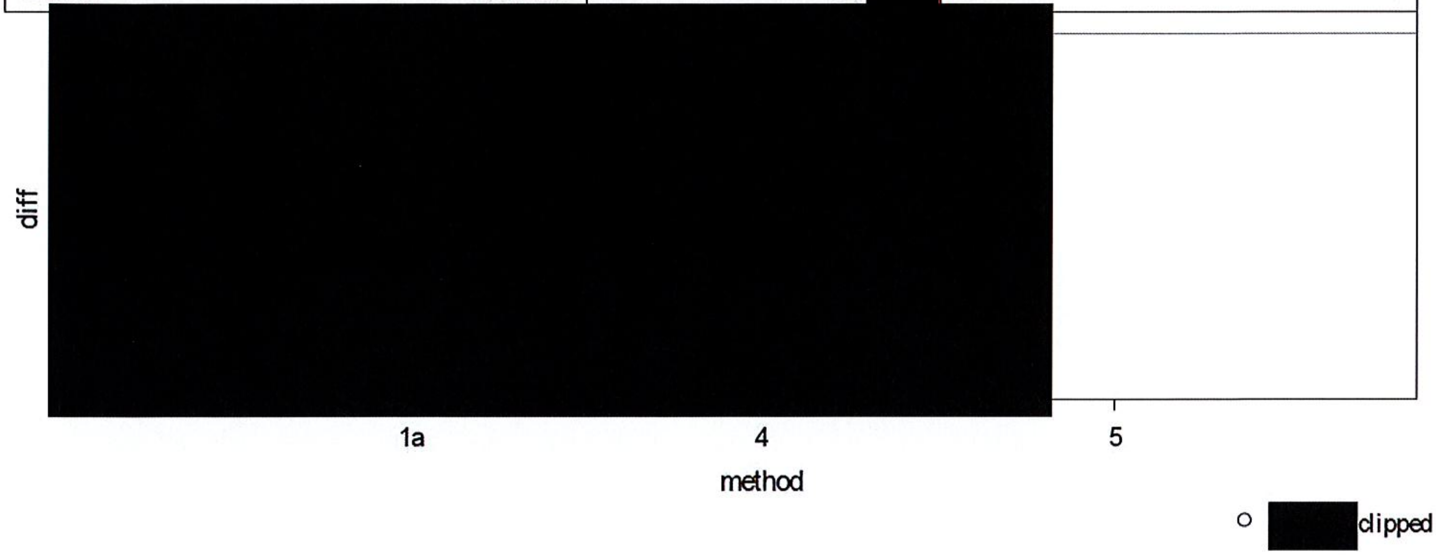
N	20	<15	<15	<15	<15	20	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	0	0	0	0	0	0	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-0.[redacted]	-2.[redacted]	-5.[redacted]	-4.[redacted]	0.[redacted]	7.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	0	0	0	0	0	0	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	0	0	0	0	0	0	0
Range	70	<15	20	20	<15	80	.
Std Dev	13.[redacted]	3.[redacted]	8.[redacted]	7.[redacted]	2.[redacted]	21.[redacted]	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=904

Statistics by GQtype			
N	6200	8600	0
Min	[redacted]	[redacted]	.
Nlow	300	0	0
Q1	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	.
Mean	-0.[redacted]	-30.[redacted]	.
Q3	[redacted]	[redacted]	.
Nhigh	300	450	0
Max	[redacted]	[redacted]	.
Nout	550	450	0
Range	500	1700	.
Std Dev	10[redacted]	31.[redacted]	.

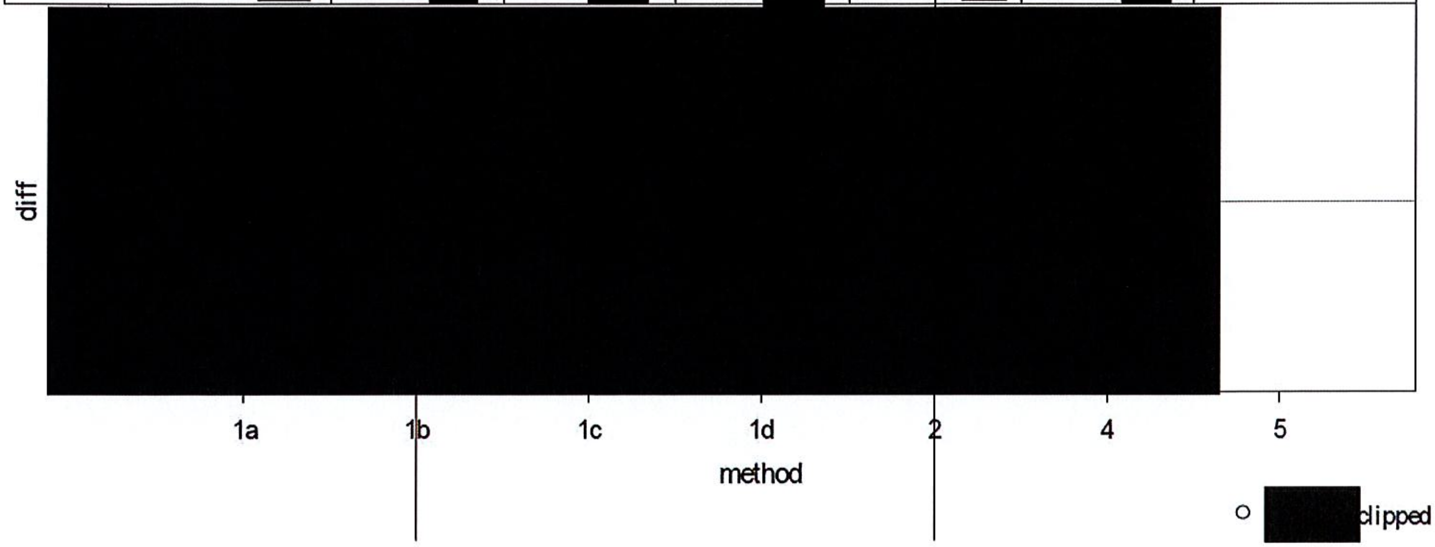

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=999

Statistics by GQtype

N	500	<15	<15	<15	<15	700	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	30	0	0	0	0	20	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-2.[redacted]	-7.[redacted]	0.[redacted]	-2.[redacted]	-2.[redacted]	12.[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	30	0	0	0	0	40	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	50	0	0	0	0	50	0
Range	400	100	30	30	160	1000	.
Std Dev	24.[redacted]	21.[redacted]	5.[redacted]	8.[redacted]	35.[redacted]	49.[redacted]	.



DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=103

Statistics by GQtype

N	550	550	550	550	550	550	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	30	30	30	30	30	30	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-2	-7	0	-7	-1	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	30	30	30	30	30	30	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	50	50	50	50	50	50	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	162	172	171	172	139	278	.

diff

[redacted]

1a 1b 1c 1d 2 4 5

method

○ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=104

Statistics by GQtype							
N	2000	2000	2000	2000	2000	2000	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	100	100	100	100	100	100	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1	-1	-1	-3	-1	6	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	100	100	100	100	100	100	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	200	200	200	200	200	200	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	118	137	114	138	145	274	.

diff

[redacted]

1a 1b 1c 1d 2 4 5

method

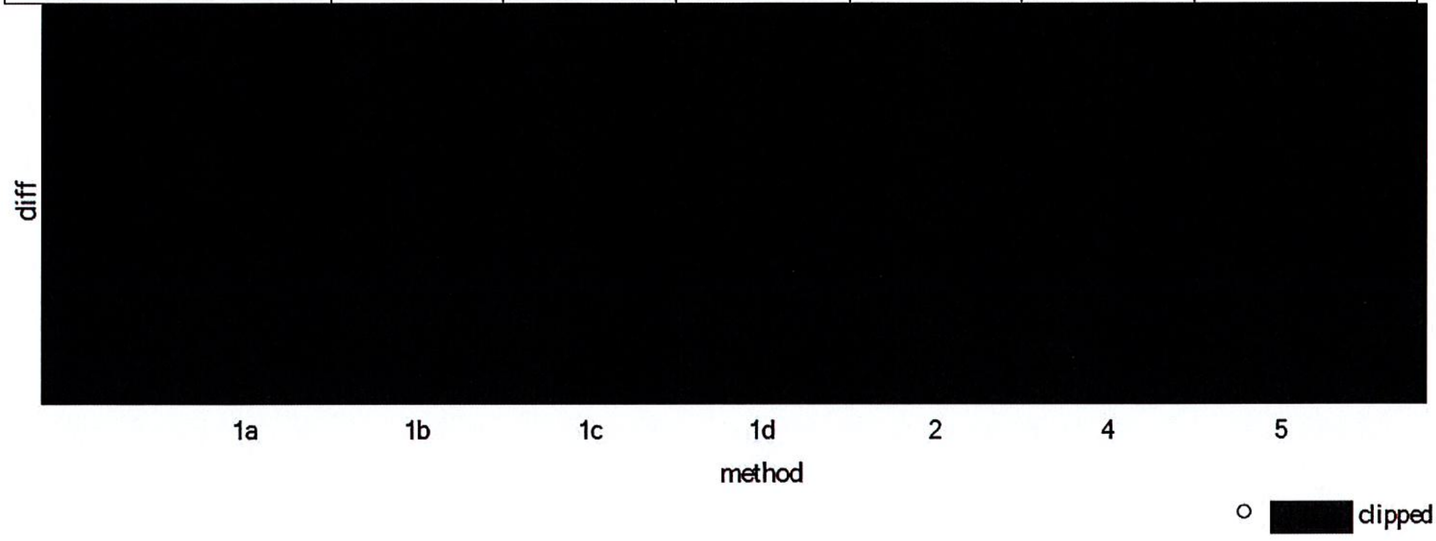
○ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=105

Statistics by GQtype

N	350	350	350	350	350	350	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	20	20	20	20	20	20	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-4	-2	2	-3	-1	2	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	20	20	20	20	20	20	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	30	30	30	30	30	30	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	65	62	64	66	47	113	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=106

Statistics by GQtype							
N	<15	<15	<15	<15	<15	<15	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	0	0	0	0	0	0	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1	-1	24	17	-1	-23	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	0	0	0	0	0	0	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	0	0	0	0	0	0	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	51	57	56	45	73	149	.

diff

[redacted]

1a 1b 1c 1d 2 4 5

method

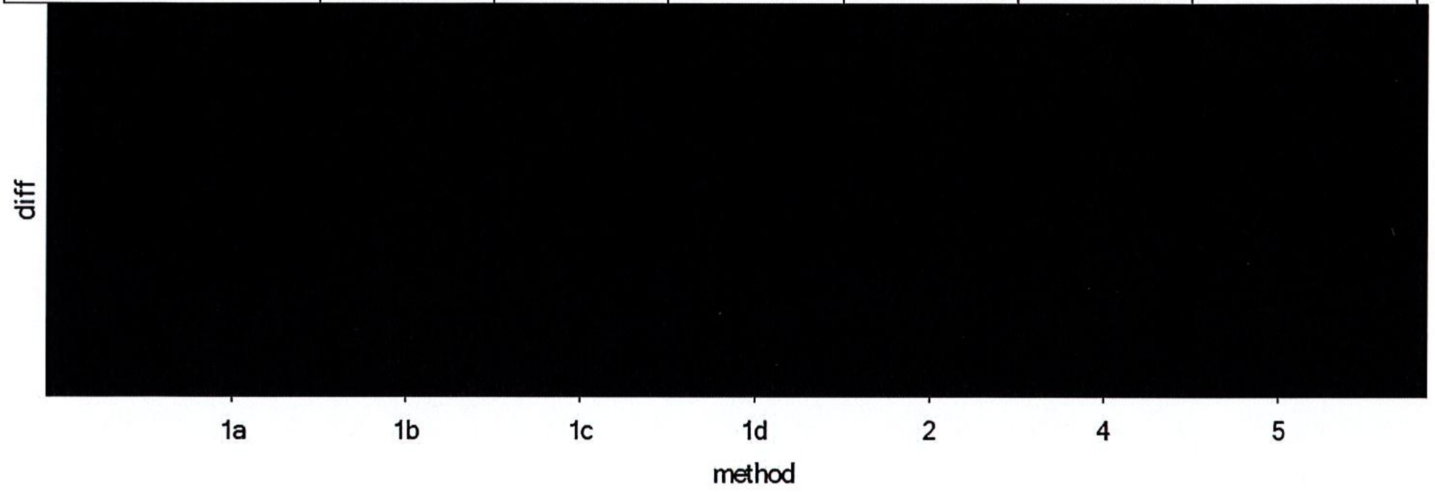
□ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=201

Statistics by GQtype

N	1400	1400	1400	1400	1400	1400	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	70	60	70	60	70	60	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0	0	-1	-1	0	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	60	70	70	60	70	70	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	150	150	150	100	150	150	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	9	9	8	9	8	13	.



[redacted]

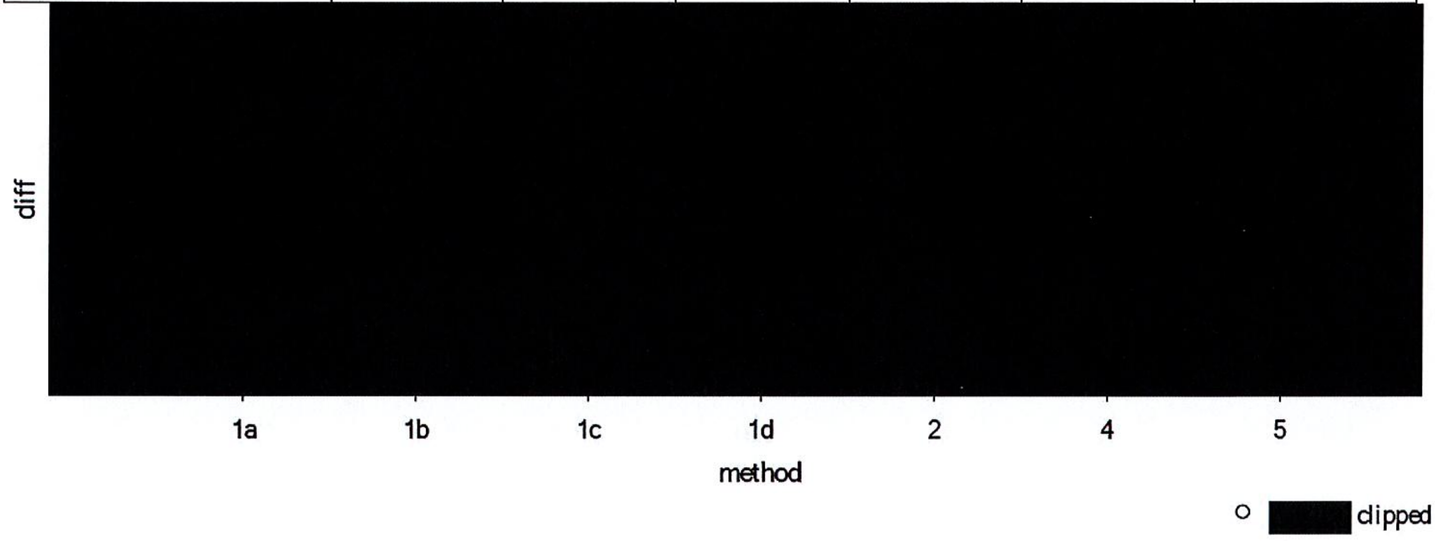
○ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=202

Statistics by GQtype

N	850	850	850	850	850	850	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	40	40	40	40	40	40	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0	–1	0	-0	-1	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	40	40	40	40	40	40	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	80	80	80	80	80	80	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	11	13	14	13	12	24	.

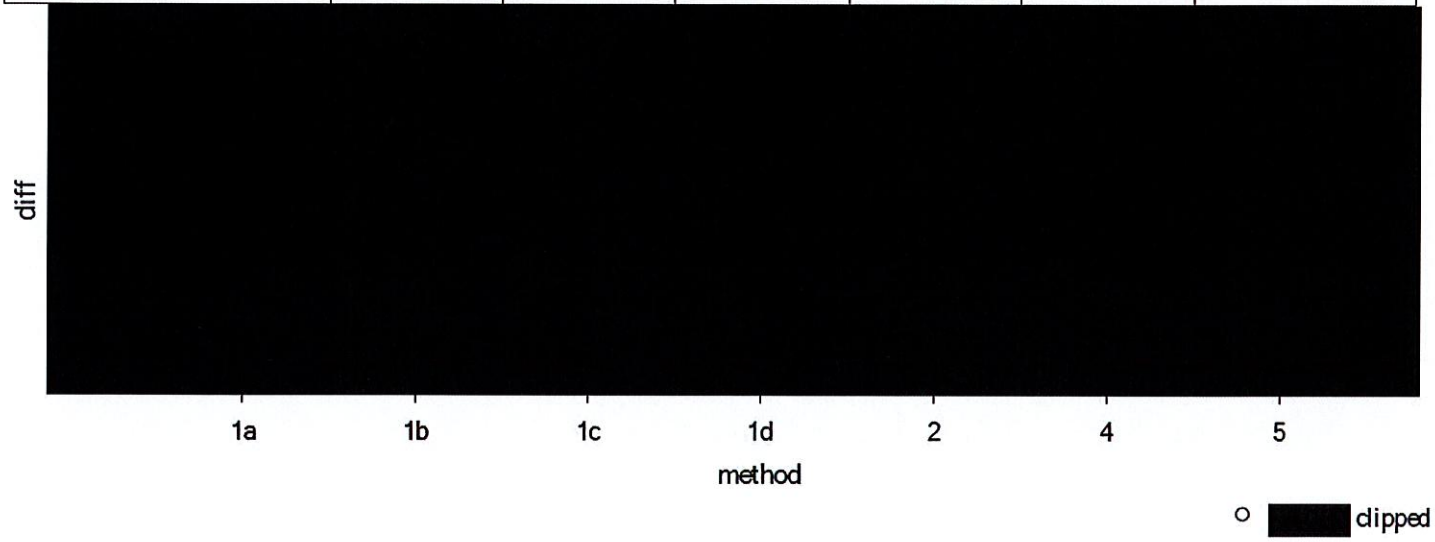

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=203

Statistics by GQtype

N	550	550	550	550	550	550	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	30	30	30	30	30	30	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-2	-3	-2	-3	-1	0	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	30	30	30	30	30	30	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	60	60	60	50	60	50	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	29	33	29	31	19	24	.

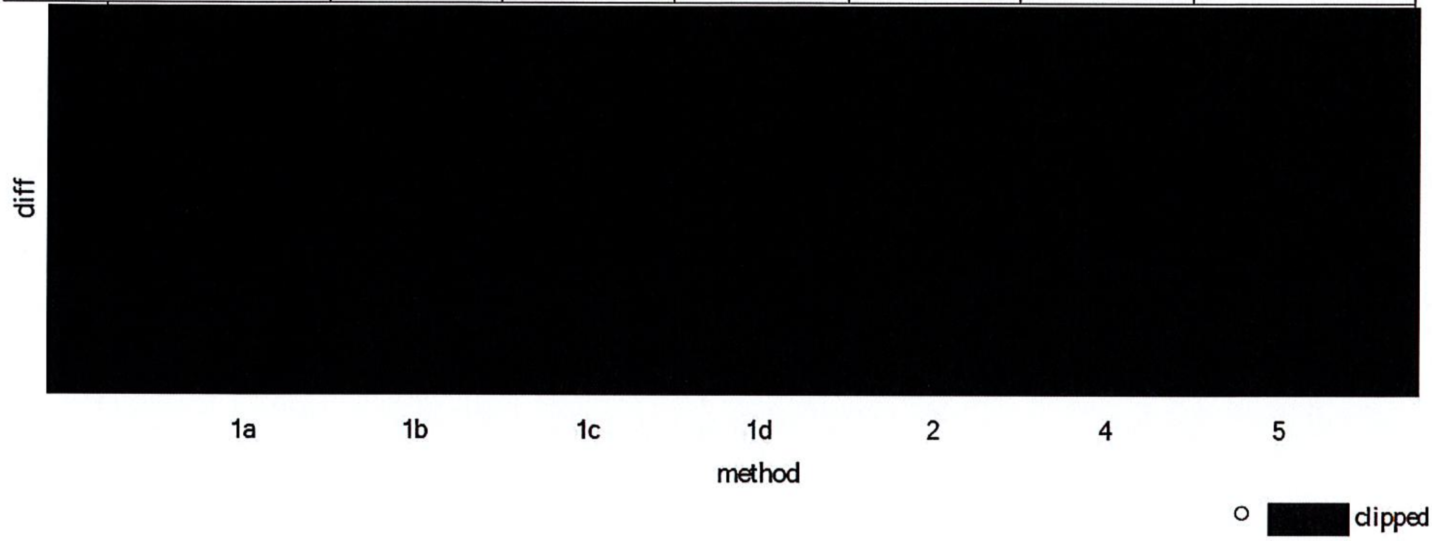

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=301

Statistics by GQtype

N	14500	14500	14500	14500	14500	14500	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	700	700	700	700	700	700	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0	0	-0	0	–1	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	700	700	700	700	700	700	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	1400	1400	1400	1400	1400	1400	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	25	28	28	28	25	50	.


[redacted]

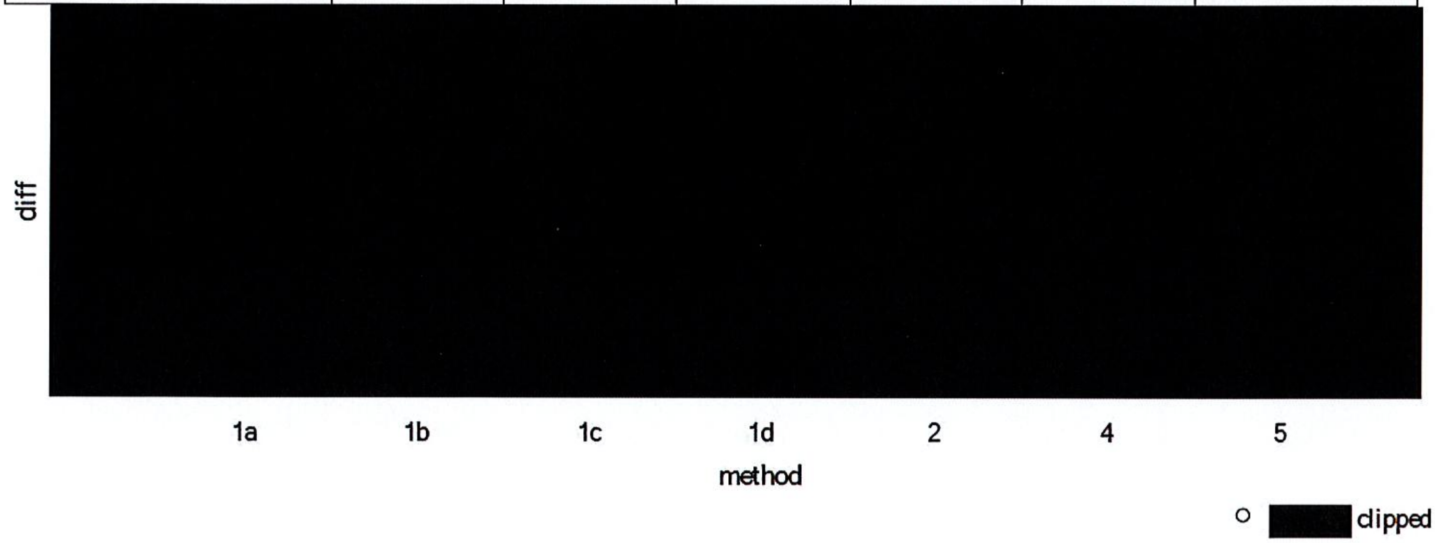
[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=401

Statistics by GQtype

N	250	250	250	250	250	250	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-5	-4	0	0	-1	3	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	20	20	20	20	20	20	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	82	83	63	79	46	101	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=402

Statistics by GQtype							
N	70	70	70	70	70	70	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-5	–3	-7	-4	0	-6	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	<15	<15	<15	<15	<15	<15	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	30	26	33	43	31	74	.

diff

[redacted]

1a 1b 1c 1d 2 4 5

method

○ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=403

Statistics by GQtype							
N	200	200	200	200	200	200	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0	0	-1	0	-1	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	20	20	20	20	20	20	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	11	9	12	9	11	25	.

diff

[redacted]

1a 1b 1c 1d 2 4 5

method

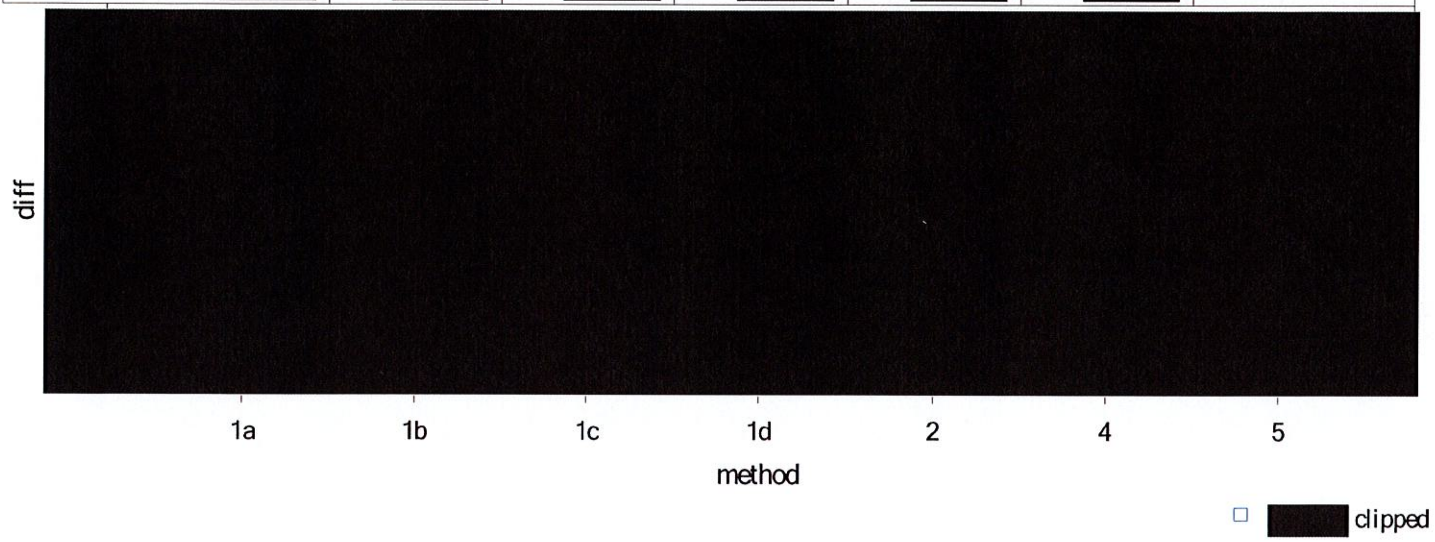
○ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=404

Statistics by GQtype

N	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.

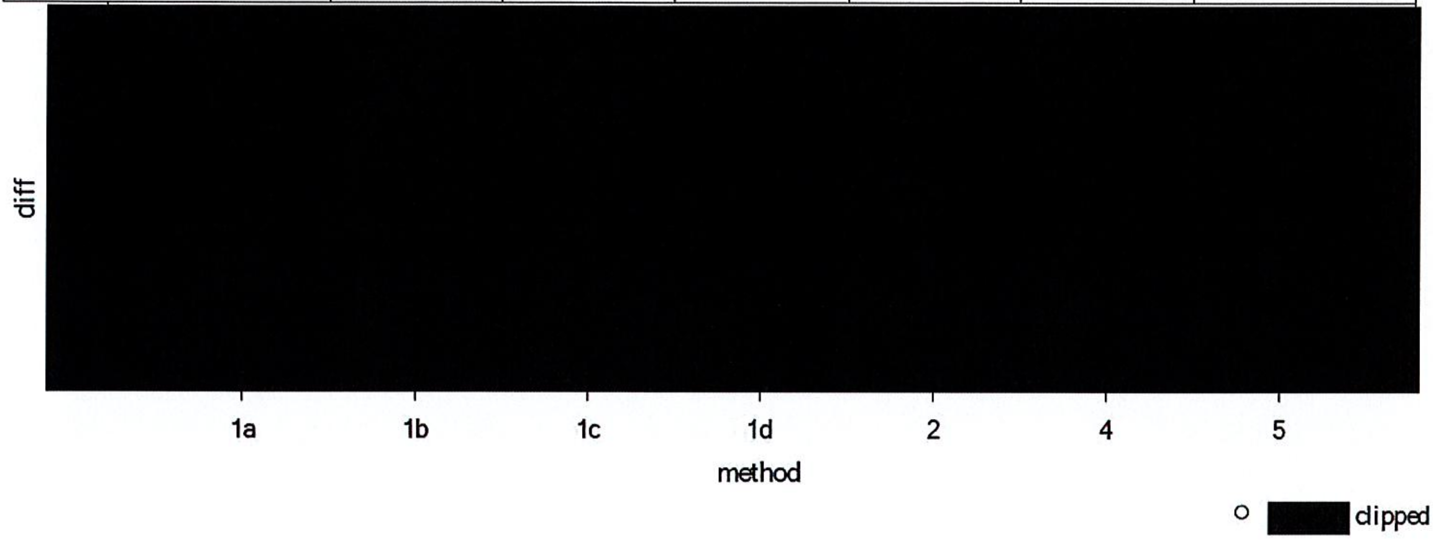

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=405

Statistics by GQtype

N	250	250	250	250	250	250	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-3	-4	0	-3	-1	-1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	20	20	20	20	20	20	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	29	31	26	30	22	49	.

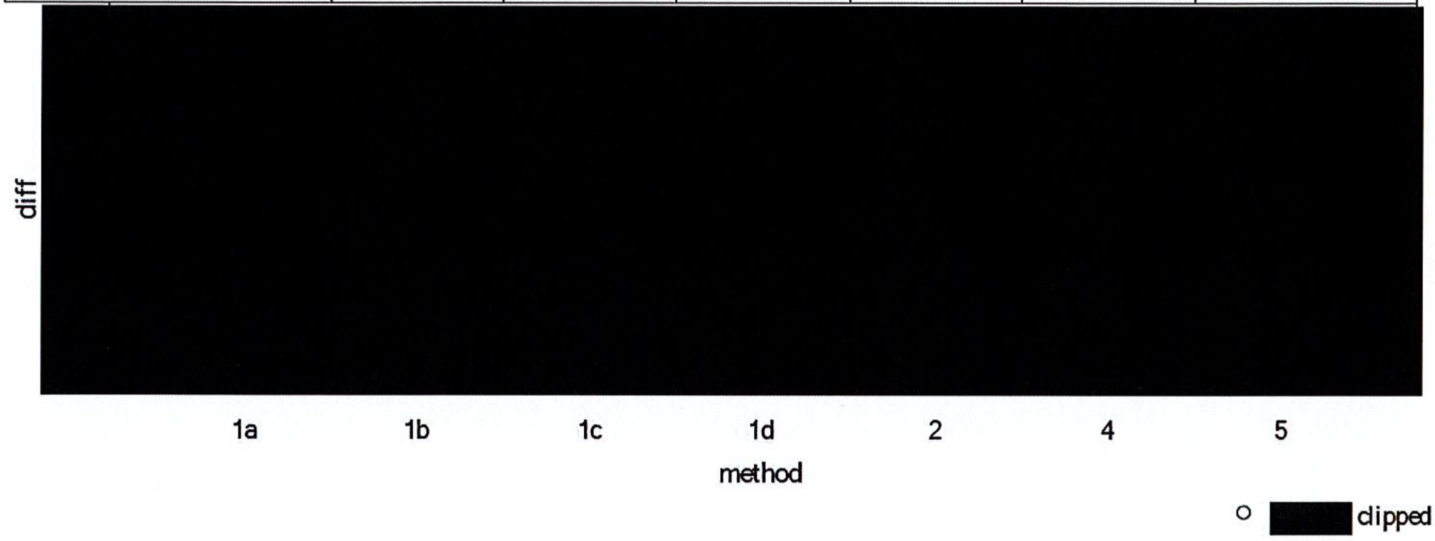

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=501

Statistics by GQtype

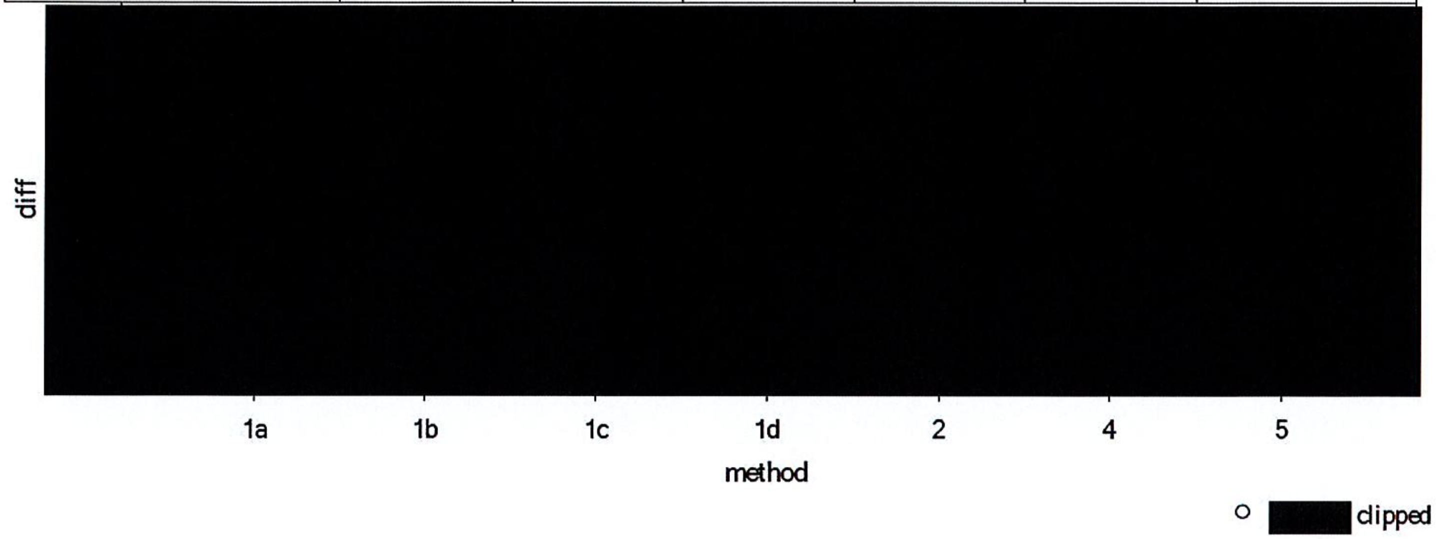
N	13000	13000	13000	13000	13000	13000	13000
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nlow	650	650	650	650	650	650	650
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	0	-1	0	-2	0	-2	0
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nhigh	650	650	650	650	650	650	650
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Nout	1300	1300	1300	1300	1300	1300	1300
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Std Dev	34	42	39	43	35	128	46


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=601

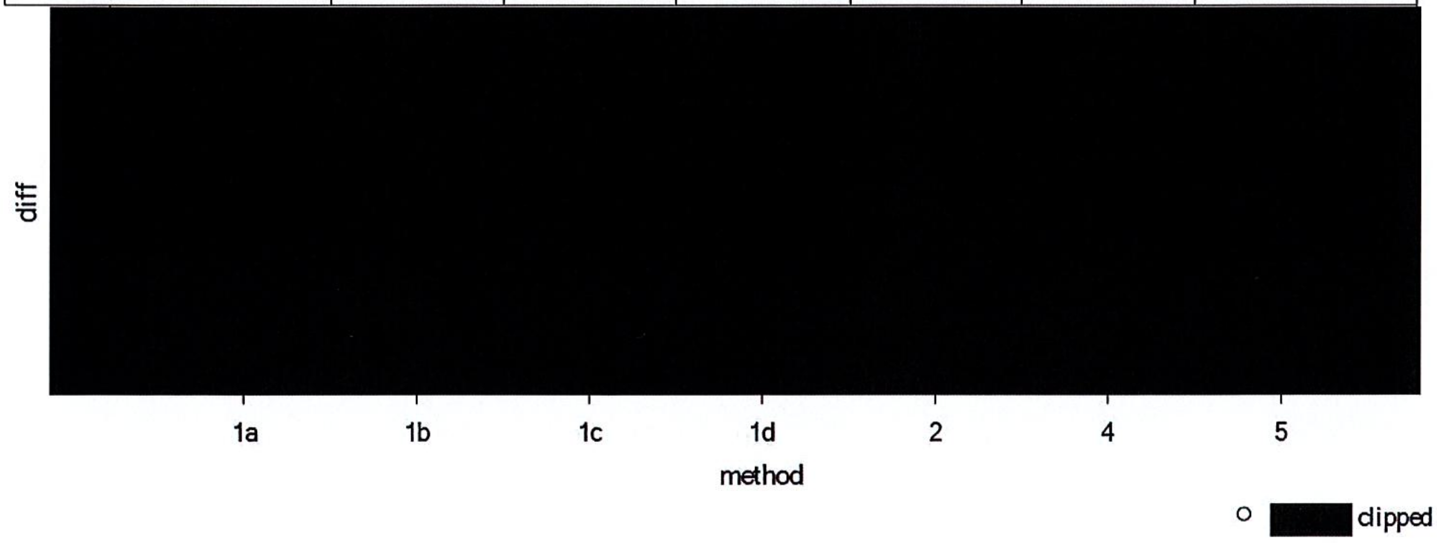
Statistics by GQtype							
N	800	800	800	800	800	800	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	40	40	40	40	40	40	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-4	-4	-5	-2	-1	-4	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	40	40	40	40	40	40	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	80	80	80	80	80	80	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	94	96	102	94	82	107	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

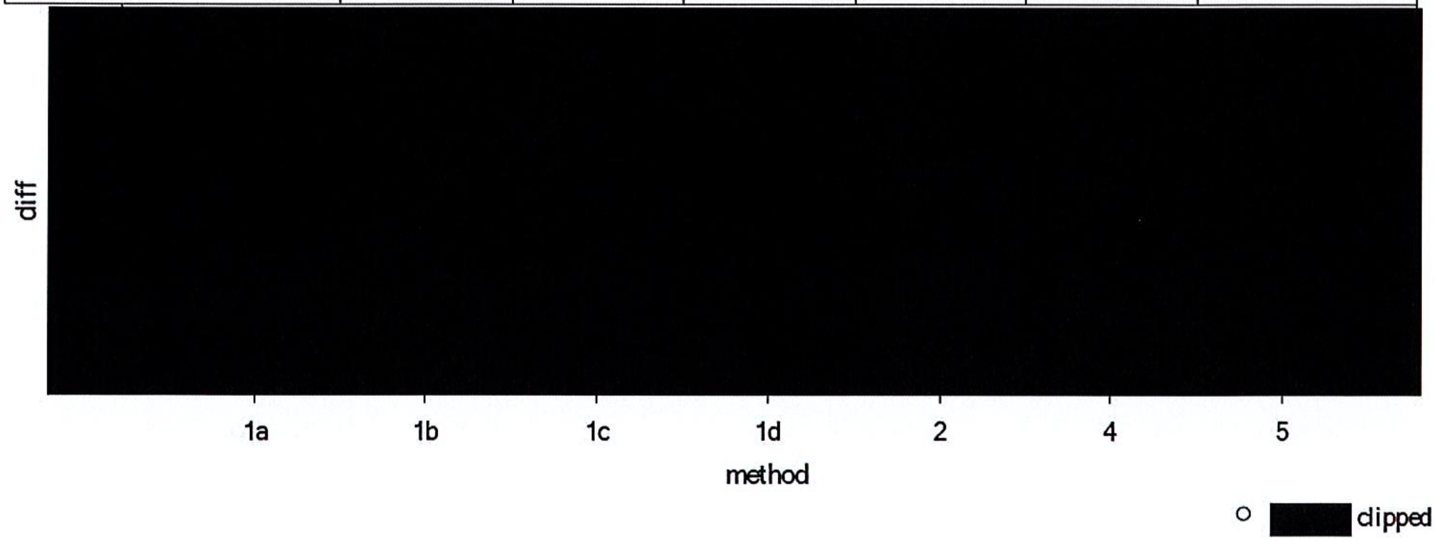
GQTYPCUR=701

Statistics by GQtype							
N	1600	1600	1600	1600	1600	1600	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	80	80	80	80	80	80	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1	-1	0	-3	0	5	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	80	80	80	80	80	80	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	150	150	150	150	150	150	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	37	36	43	37	33	59	.



GQTYPCUR=702

Statistics by GQtype							
N	700	700	700	700	700	700	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	40	40	40	40	40	30	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-3	-7	-2	-13	0	3	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	40	40	40	40	40	40	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	70	70	70	70	70	70	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	63	59	82	62	62	92	.

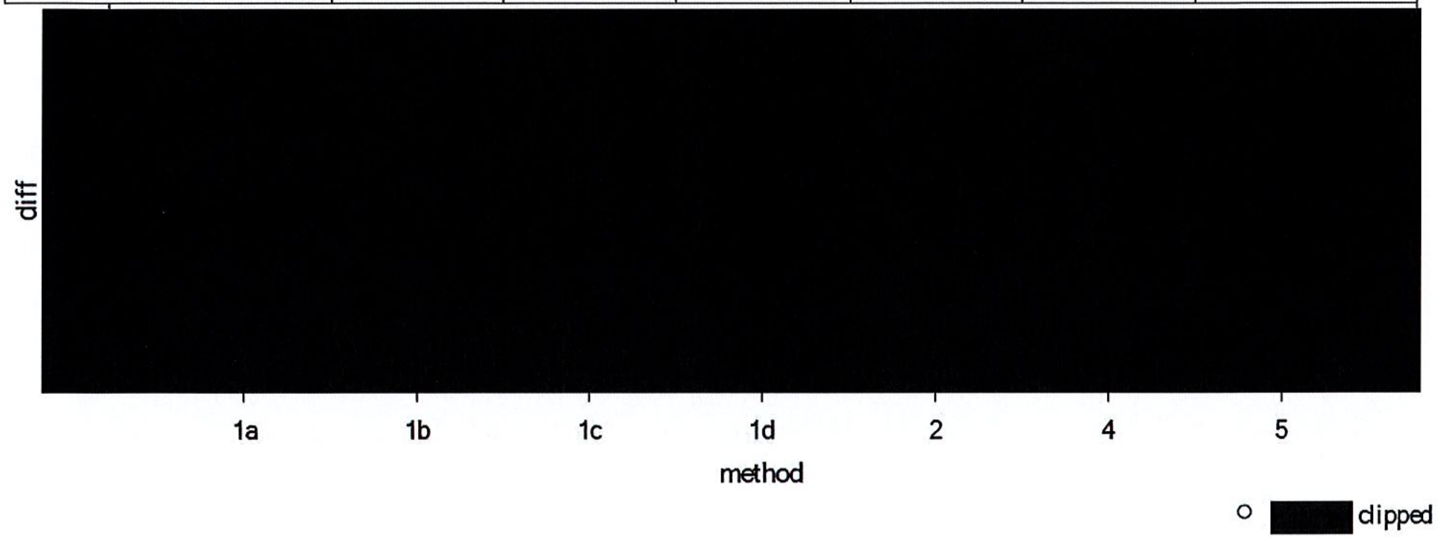

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=704

Statistics by GQtype

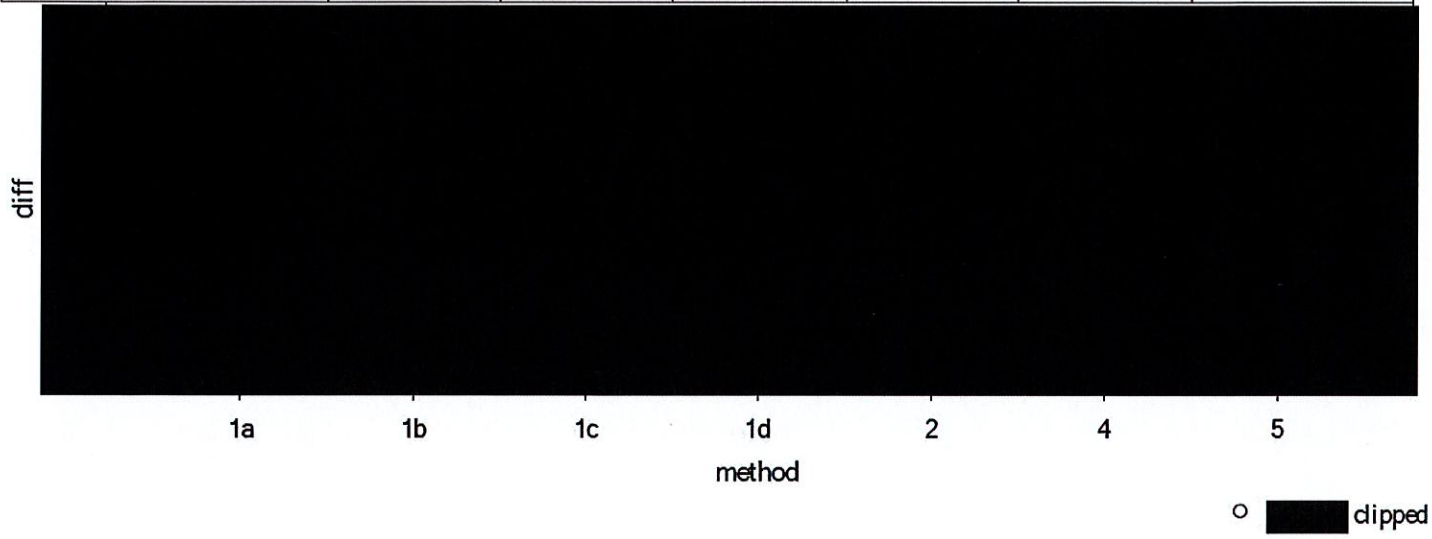
N	30	30	30	30	30	30	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	2	-20	2	-21	0	6	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	<15	<15	<15	<15	<15	<15	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	21	116	17	116	18	24	.


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=706

Statistics by GQtype							
N	200	200	200	200	200	200	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	<15	<15	<15	<15	<15	<15	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1	-5	-3	-1	-1	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	<15	<15	<15	<15	<15	<15	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	20	20	20	20	20	20	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	18	22	31	17	22	19	.

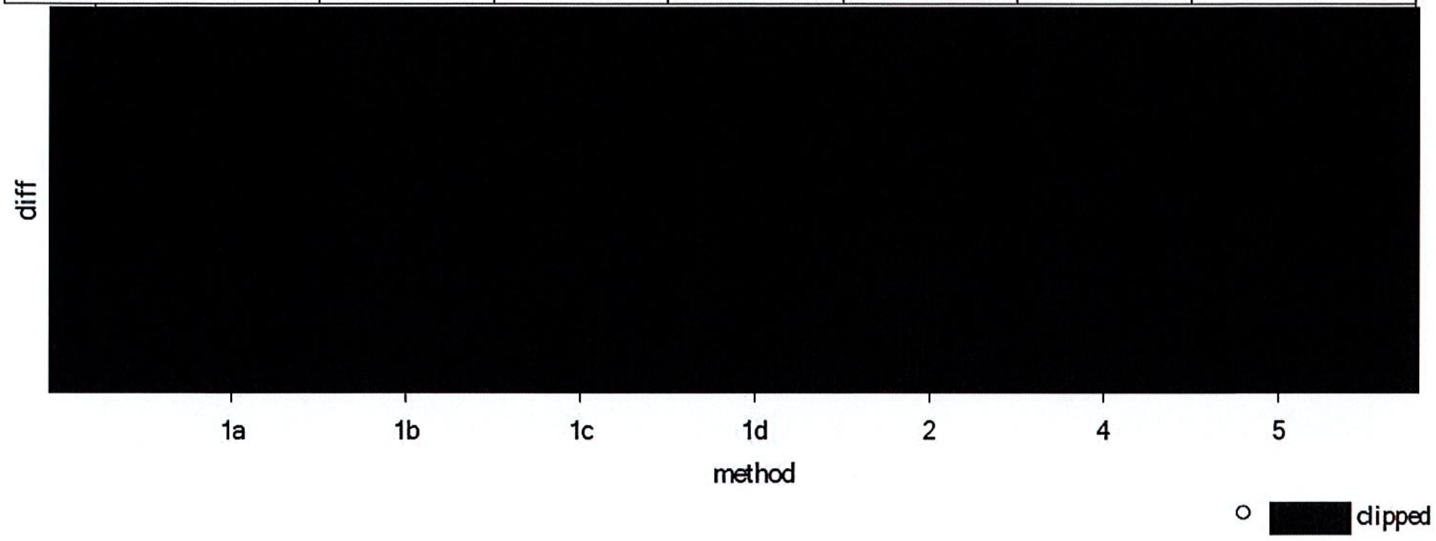

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=801

Statistics by GQtype

N	21000	21000	21000	21000	21000	21000	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	850	1000	950	1000	900	850	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	0	0	0	0	-1	1	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	950	800	900	750	850	950	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	1800	1800	1800	1800	1800	1800	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	5	6	6	6	6	12	.

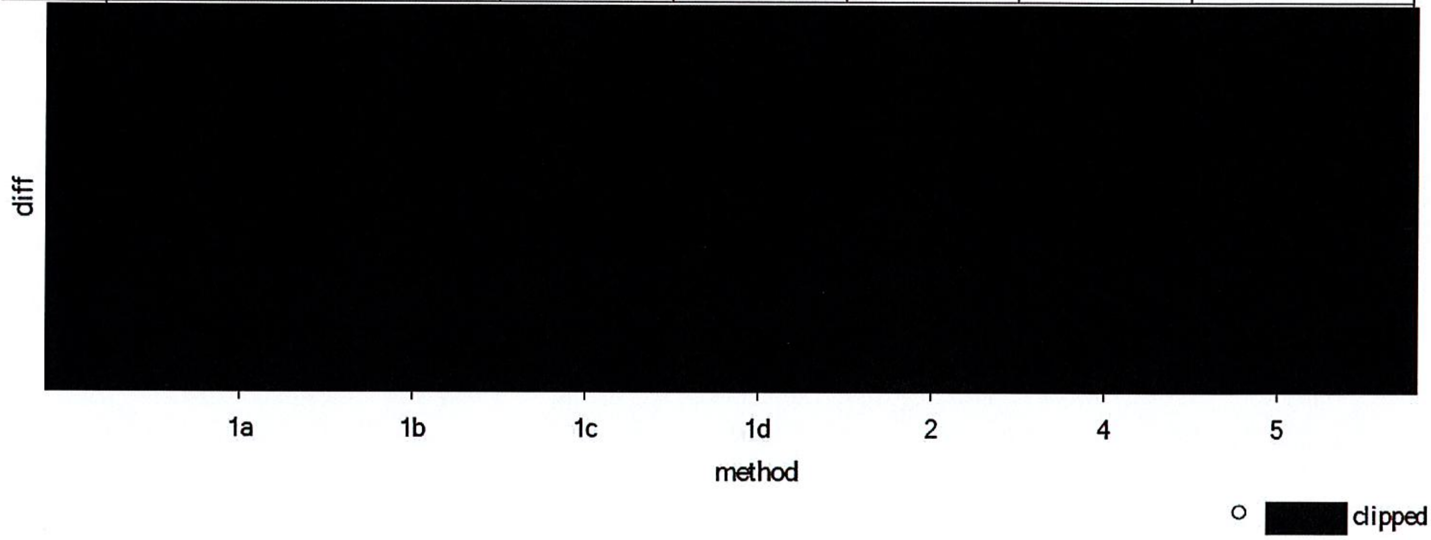

[redacted]

Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=802

Statistics by GQtype

N	3500	3500	3500	3500	3500	3500	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	150	150	150	200	150	150	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1	-1	0	-1	-1	2	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	150	150	150	150	150	150	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	350	350	350	350	350	350	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	12	13	13	13	12	29	.


[redacted]

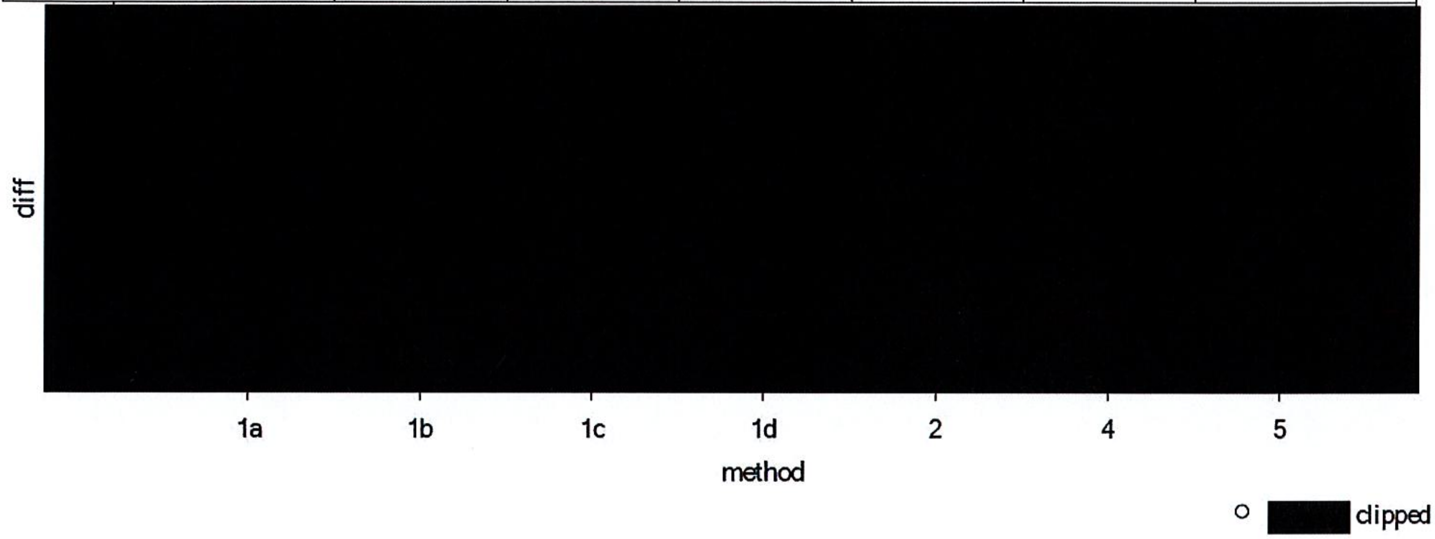
[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=901

Statistics by GQtype

N	2200	2200	2200	2200	2200	2200	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	100	100	100	100	100	100	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-1	-1	-1	–1	0	3	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	100	100	100	100	100	100	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	200	200	200	200	200	200	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	15	15	19	15	14	27	.


[redacted]

GQTYPCUR=903

Statistics by GQtype

N	<15	<15	<15	<15	<15	<15	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	0	0	0	0	0	0	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	6	-2	-6	-4	0	17	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	0	0	0	0	0	0	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	0	0	0	0	0	0	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	19	3	9	7	2	36	.

diff

[redacted]

1a 1b 1c 1d 2 4 5

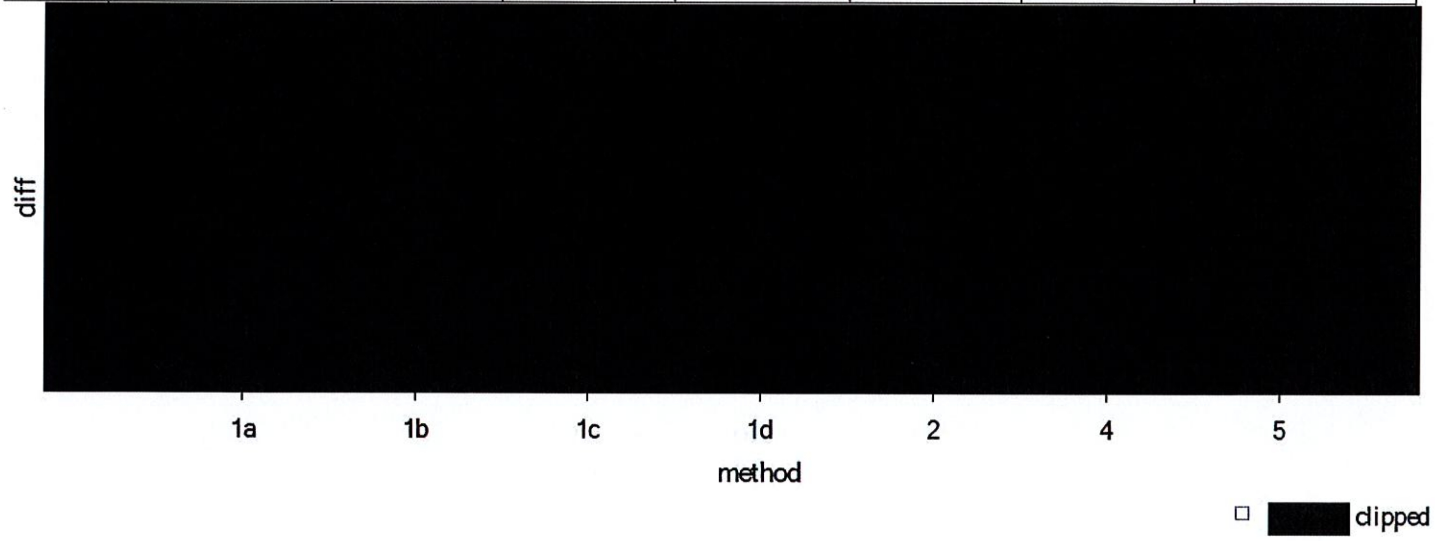
method

□ [redacted] clipped

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=999

Statistics by GQtype							
N	<15	<15	<15	<15	<15	<15	0
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nlow	0	0	0	0	0	0	0
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Mean	-17	-8	0	-3	-2	14	.
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nhigh	0	0	0	0	0	0	0
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Nout	0	0	0	0	0	0	0
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	.
Std Dev	51	22	6	8	36	34	.


[redacted]

Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=103

Overall Statistics					
Min	[redacted]	Mean	3.[redacted]	Max	[redacted]
Pooled Std Dev	264.[redacted]				

Statistics by GQtype						
N	550	550	550	550	550	550
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-8.[redacted]	28.[redacted]	-8.[redacted]	-0.[redacted]	12.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	3100	3500	6500	3500	4600	6200
Std Dev	195.[redacted]	219.[redacted]	313.[redacted]	218.[redacted]	234.[redacted]	363.[redacted]

diff

[redacted]

1a 1b 1c 1d 2 4

method

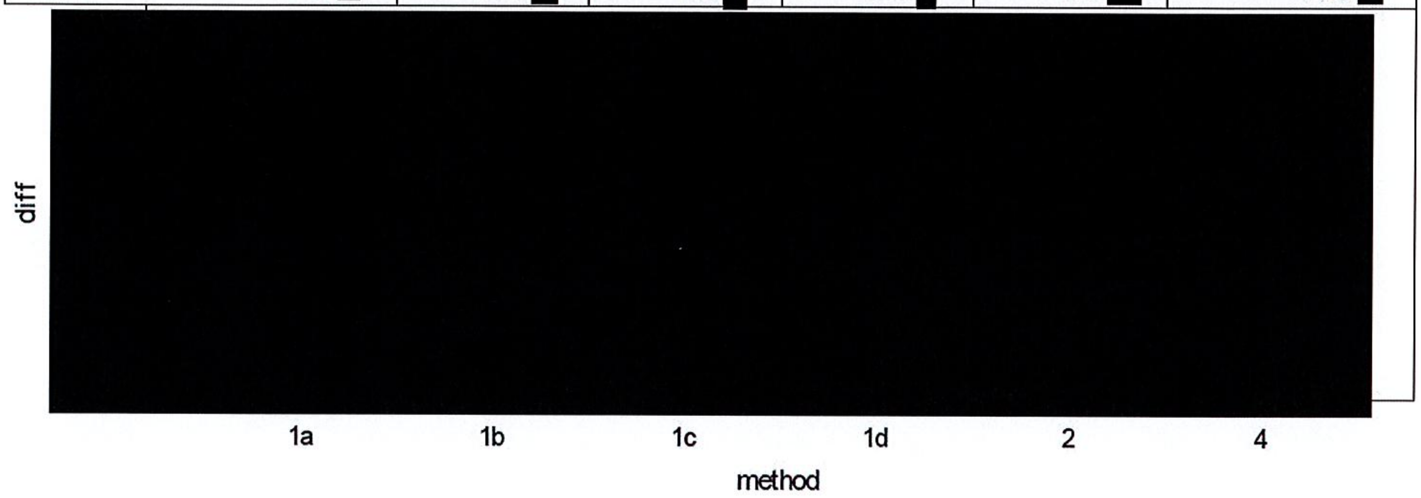
DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=104

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	205.[redacted]				

Statistics by GQtype

N	2100	2100	2100	2100	2100	2100
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean ☺	-1.[redacted]	-8.[redacted]	11.[redacted]	-12.[redacted]	-0.[redacted]	5.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	3000	8200	3100	10500	3800	4200
Std Dev	139.[redacted]	220.[redacted]	155.[redacted]	250.[redacted]	137.[redacted]	283.[redacted]

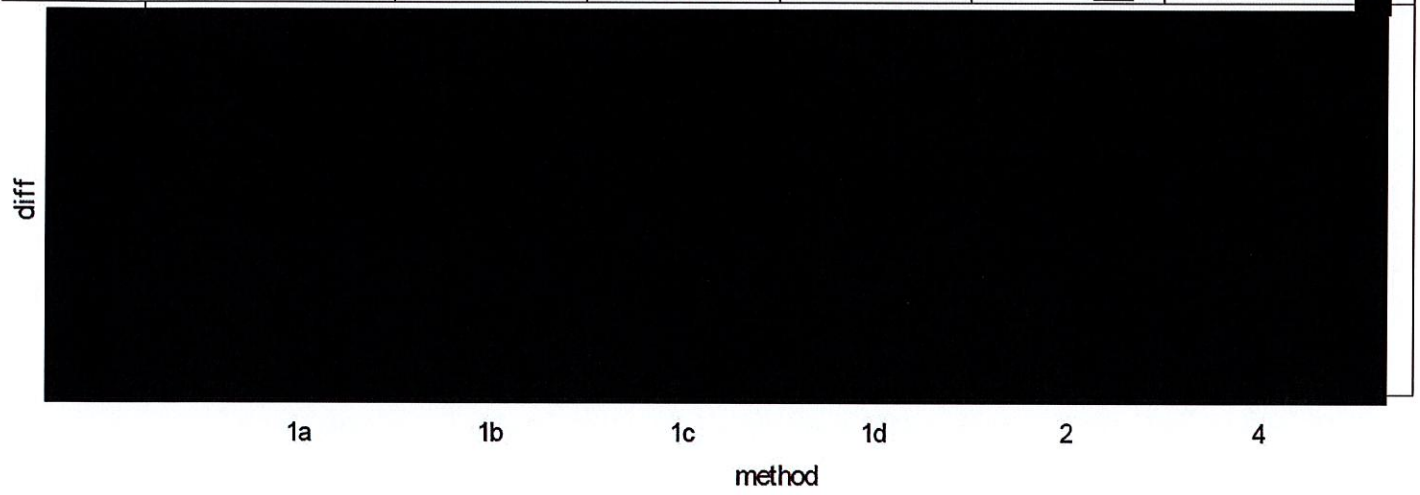

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=105

Overall Statistics					
Min	[redacted]	Mean	-74.[redacted]	Max	[redacted]
Pooled Std Dev	2398				

Statistics by GQtype						
N	350	350	350	350	350	350
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-232.[redacted]	6.[redacted]	-227.[redacted]	-0.[redacted]	4.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	1100	77000	1300	74500	800	1800
Std Dev	74.[redacted]	4223	90.[redacted]	4079	62.[redacted]	137.[redacted]


[redacted]

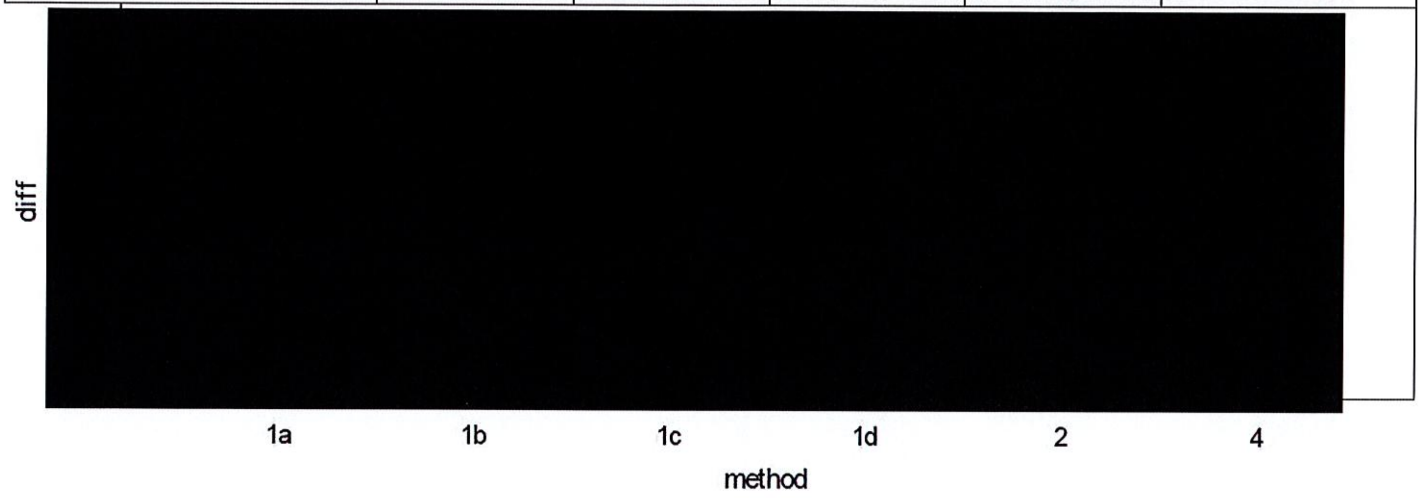
DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=106

Overall Statistics					
Min	[redacted]	Mean	8.[redacted]	Max	[redacted]
Pooled Std Dev	79.[redacted]				

Statistics by GQtype

N	<15	<15	<15	<15	<15	<15
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	2	1	26	17	5	-2
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	150	150	150	100	200	350
Std Dev	51	56	55	44	68	148


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=201

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	11.[redacted]				

Statistics by GQtype						
N	1500	1500	1500	1500	1500	1500
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	200	200	450	200	200	250
Std Dev	9.[redacted]	10.[redacted]	13.[redacted]	10.[redacted]	10.[redacted]	15.[redacted]

diff

[redacted]

1a 1b 1c 1d 2 4

method

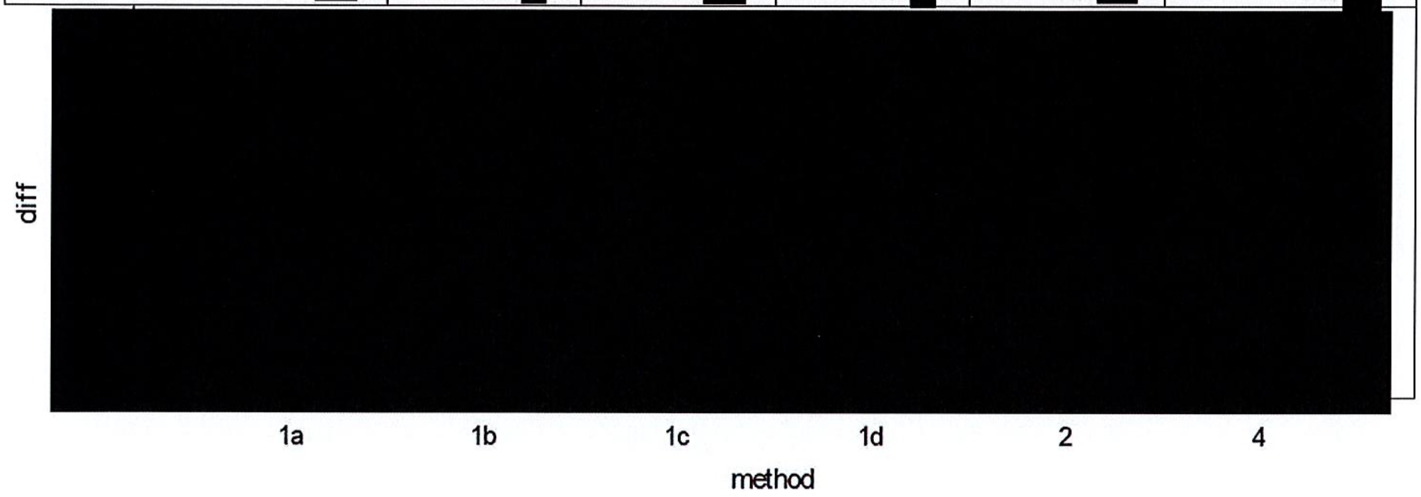
DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=202

Overall Statistics					
Min	[redacted]	Mean	-2.[redacted]	Max	[redacted]
Pooled Std Dev	92.[redacted]				

Statistics by GQtype						
N	850	850	850	850	850	850
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-7.[redacted]	-0.[redacted]	-6.[redacted]	-0.[redacted]	1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	200	4700	300	4700	500	300
Std Dev	12.[redacted]	158.[redacted]	17.[redacted]	157.[redacted]	19.[redacted]	25.[redacted]

[redacted]

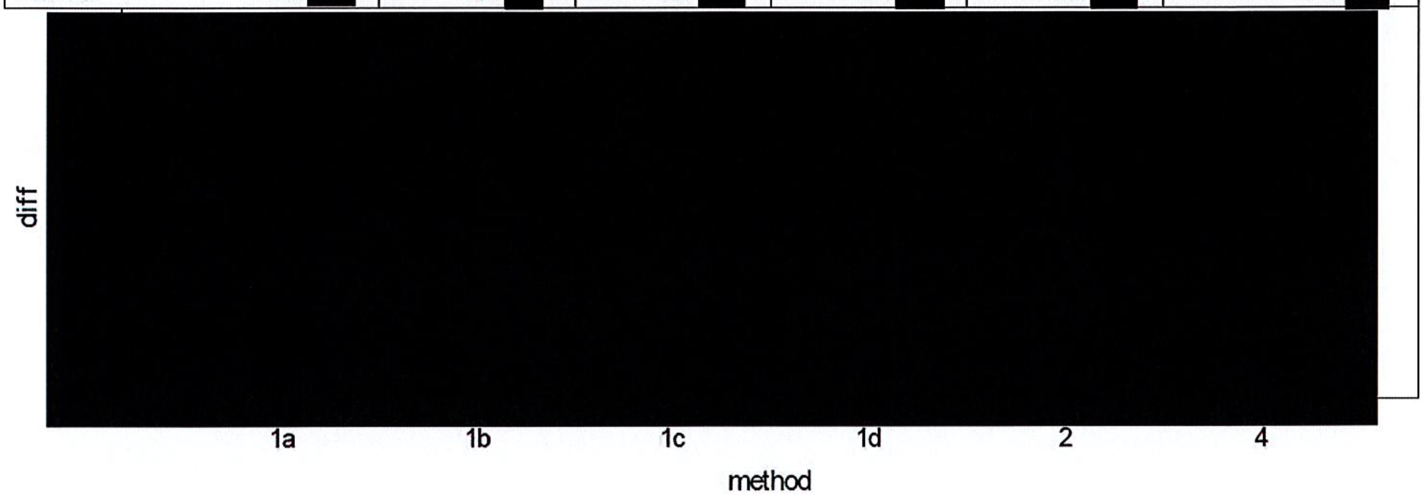


DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=203

Overall Statistics					
Min	[redacted]	Mean	-1.[redacted]	Max	[redacted]
Pooled Std Dev	31.[redacted]				

Statistics by GQtype						
N	600	600	600	600	600	600
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-2.[redacted]	-3.[redacted]	-2.[redacted]	-3.[redacted]	-0.[redacted]	0.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	550	550	700	550	300	300
Std Dev	32.[redacted]	32.[redacted]	39.[redacted]	31.[redacted]	22.[redacted]	27.[redacted]



[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=301

Overall Statistics					
Min	[redacted]	Mean	-3.[redacted]	Max	[redacted]
Pooled Std Dev	386.[redacted]				

Statistics by GQtype						
N	14500	14500	14500	14500	14500	14500
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-11.[redacted]	0[redacted]	-11.[redacted]	-0.[redacted]	0.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	1100	75000	1400	75000	1200	1400
Std Dev	25.[redacted]	667.[redacted]	33.[redacted]	668.[redacted]	28.[redacted]	51.[redacted]

diff

[redacted]

1a 1b 1c 1d 2 4

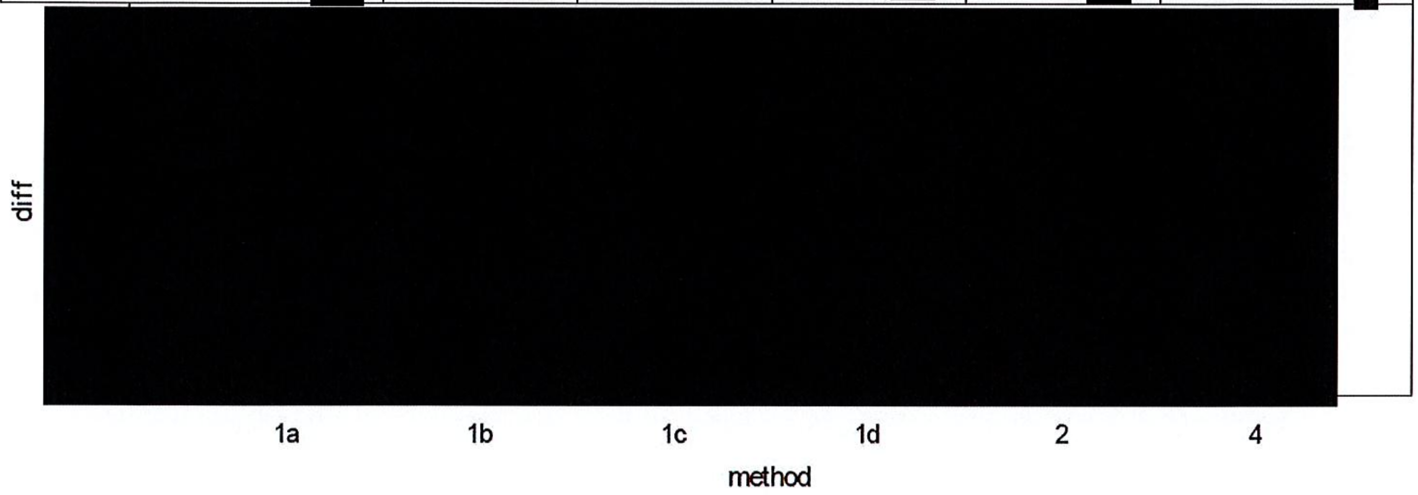
method

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=401

Overall Statistics					
Min	[redacted]	Mean	1.[redacted]	Max	[redacted]
Pooled Std Dev	93.[redacted]				

Statistics by GQtype						
N	250	250	250	250	250	250
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-3.[redacted]	-3.[redacted]	10.[redacted]	-0.[redacted]	-0.[redacted]	6.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	1400	1400	1600	1400	700	1600
Std Dev	77.[redacted]	79[redacted]	107.[redacted]	80.[redacted]	56.[redacted]	135.[redacted]

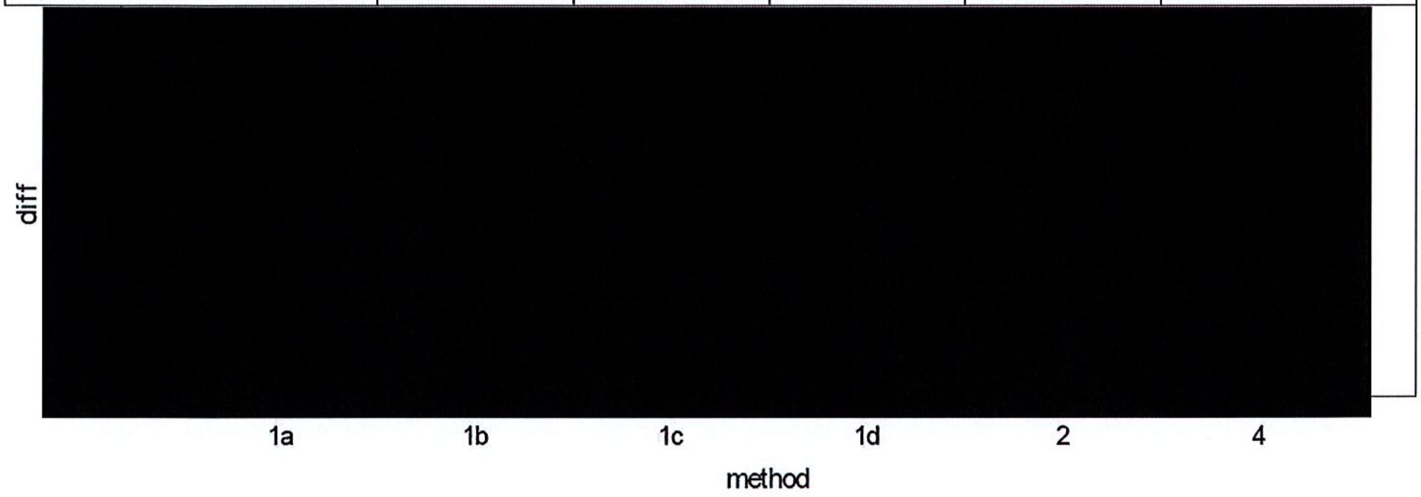

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=402

Overall Statistics					
Min	[redacted]	Mean	-4.[redacted]	Max	[redacted]
Pooled Std Dev	52				

Statistics by GQtype						
N	70	70	70	70	70	70
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	0.[redacted]	-3	-10	-4	-0.[redacted]	-6
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	150	200	650	400	300	600
Std Dev	20	30	70	40	40	80

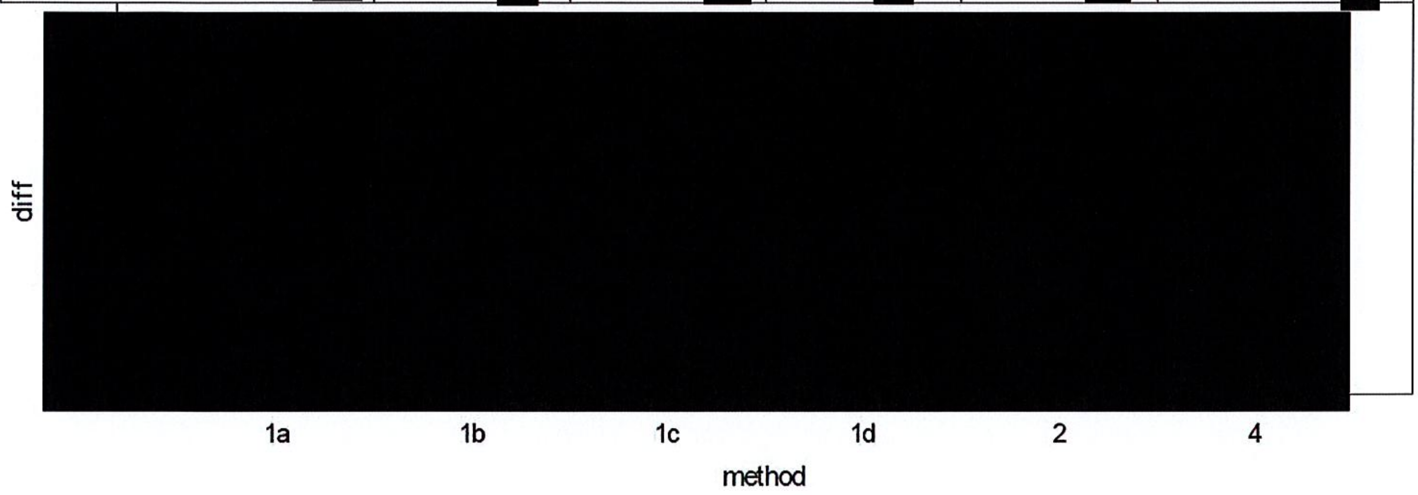

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=403

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	15.[redacted]				

Statistics by GQtype						
N	200	200	200	200	200	200
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	150	100	250	100	150	300
Std Dev	10.[redacted]	10.[redacted]	19.[redacted]	9.[redacted]	10.[redacted]	26.[redacted]

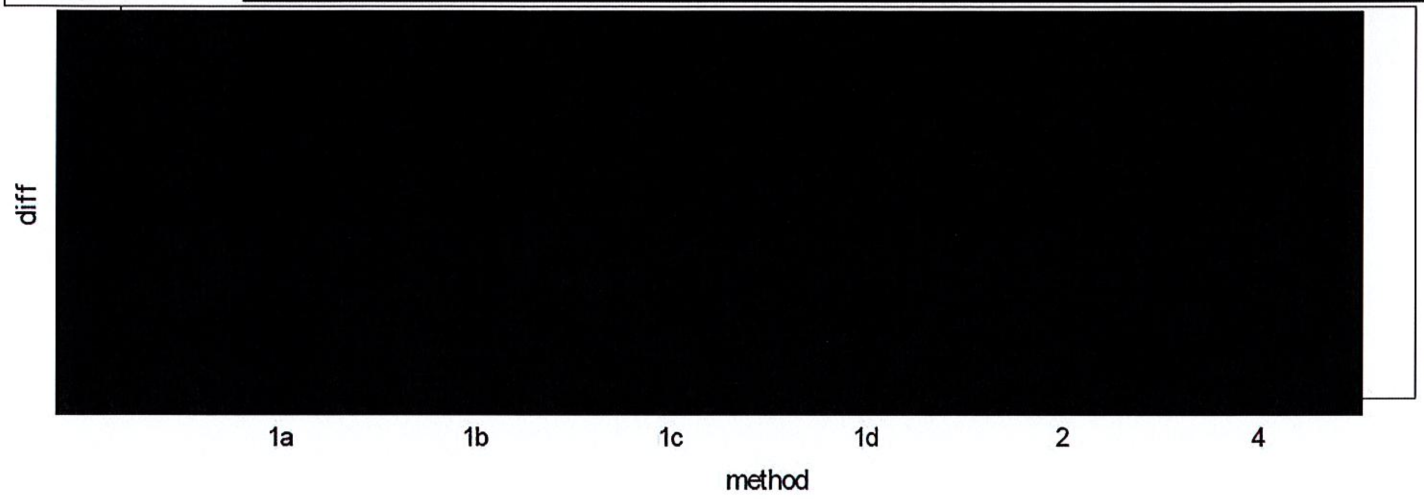

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=404

Overall Statistics					
Min	[redacted]	Mean	-14	Max	[redacted]
Pooled Std Dev	36				

Statistics by GQtype						
N	<15	<15	<15	<15	<15	<15
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Std Dev	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=405

Overall Statistics					
Min	[redacted]	Mean	-2.[redacted]	Max	[redacted]
Pooled Std Dev	34.[redacted]				

Statistics by GQtype						
N	250	250	250	250	250	250
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-3.[redacted]	-5.[redacted]	0.[redacted]	-5.[redacted]	-0.[redacted]	0[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	250	250	500	250	250	550
Std Dev	27.[redacted]	30.[redacted]	38.[redacted]	29.[redacted]	24.[redacted]	48.[redacted]

diff

[redacted]

1a 1b 1c 1d 2 4

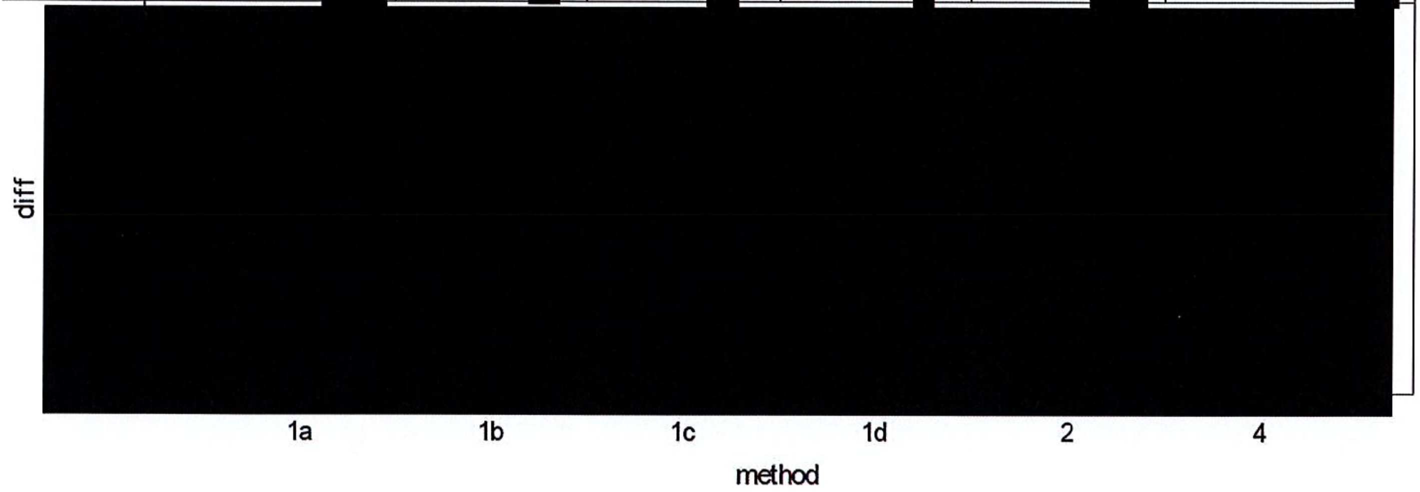
method

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=501

Overall Statistics					
Min	[redacted]	Mean	-2.[redacted]	Max	[redacted]
Pooled Std Dev	236.[redacted]				

Statistics by GQtype						
N	13000	13000	13000	13000	13000	13000
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-6.[redacted]	1.[redacted]	-7.[redacted]	-0.[redacted]	-1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	1200	41000	3500	42500	1600	3000
Std Dev	34.[redacted]	387.[redacted]	55.[redacted]	401.[redacted]	40.[redacted]	131.[redacted]


[redacted]

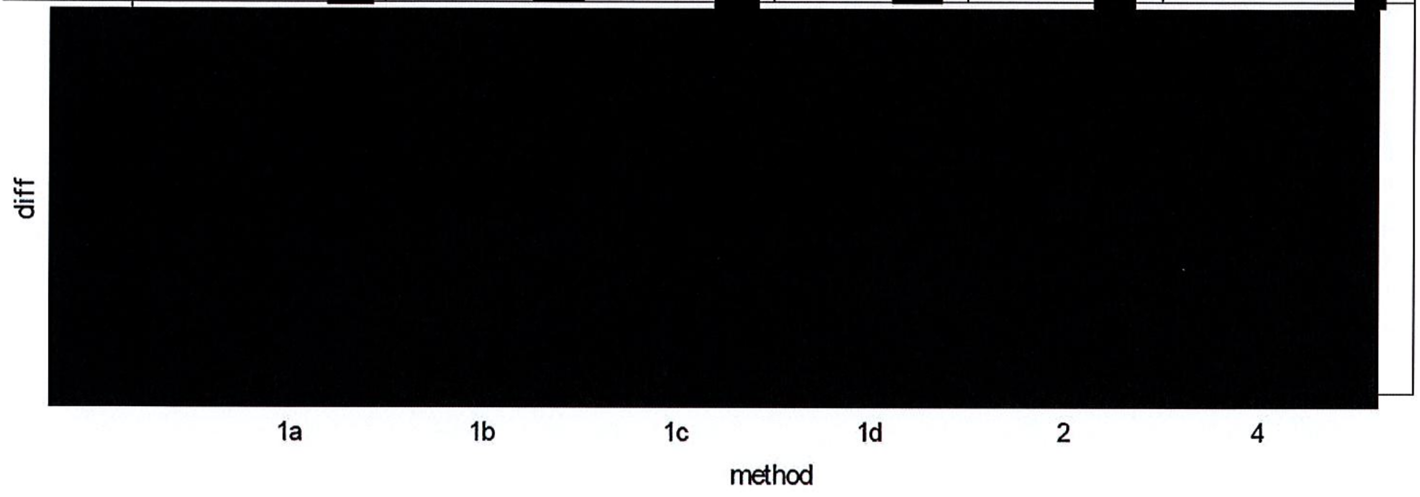
DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=601

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	106.[redacted]				

Statistics by GQtype

N	850	850	850	850	850	850
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-1.[redacted]	-4.[redacted]	3.[redacted]	-0[redacted]	0.[redacted]	0.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	1400	1600	2000	1500	1100	1300
Std Dev	93.[redacted]	97.[redacted]	129.[redacted]	94[redacted]	90.[redacted]	127.[redacted]

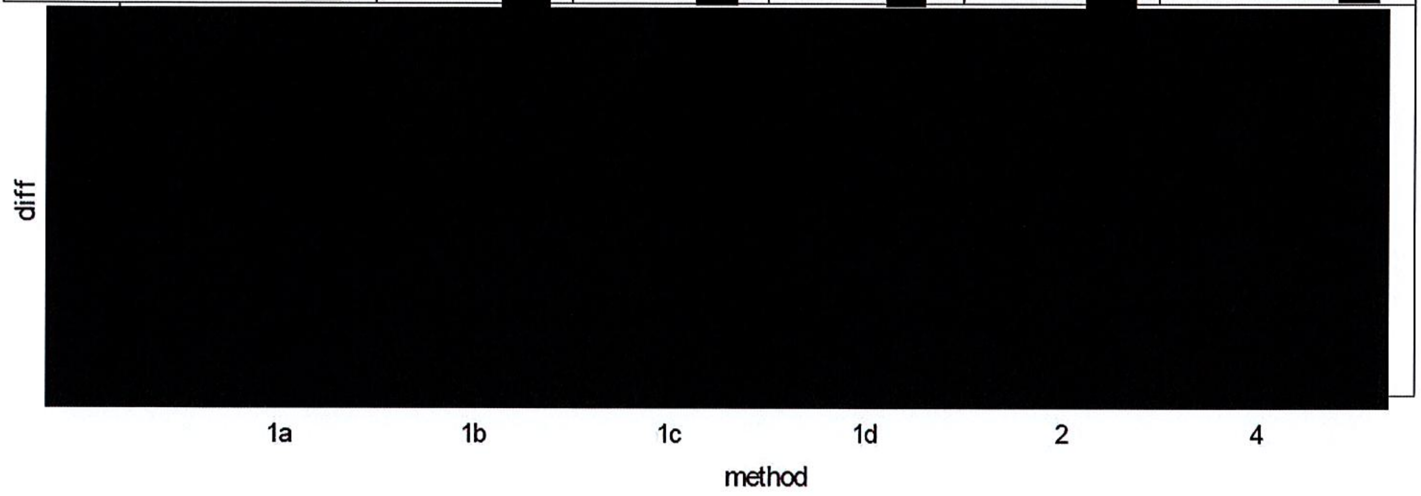

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=701

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	41.[redacted]				

Statistics by GQtype						
N	1600	1600	1600	1600	1600	1600
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-1.[redacted]	-1.[redacted]	-0.[redacted]	-3.[redacted]	-0.[redacted]	4.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	900	800	750	800	600	1200
Std Dev	36.[redacted]	35.[redacted]	42.[redacted]	36[redacted]	34.[redacted]	58[redacted]

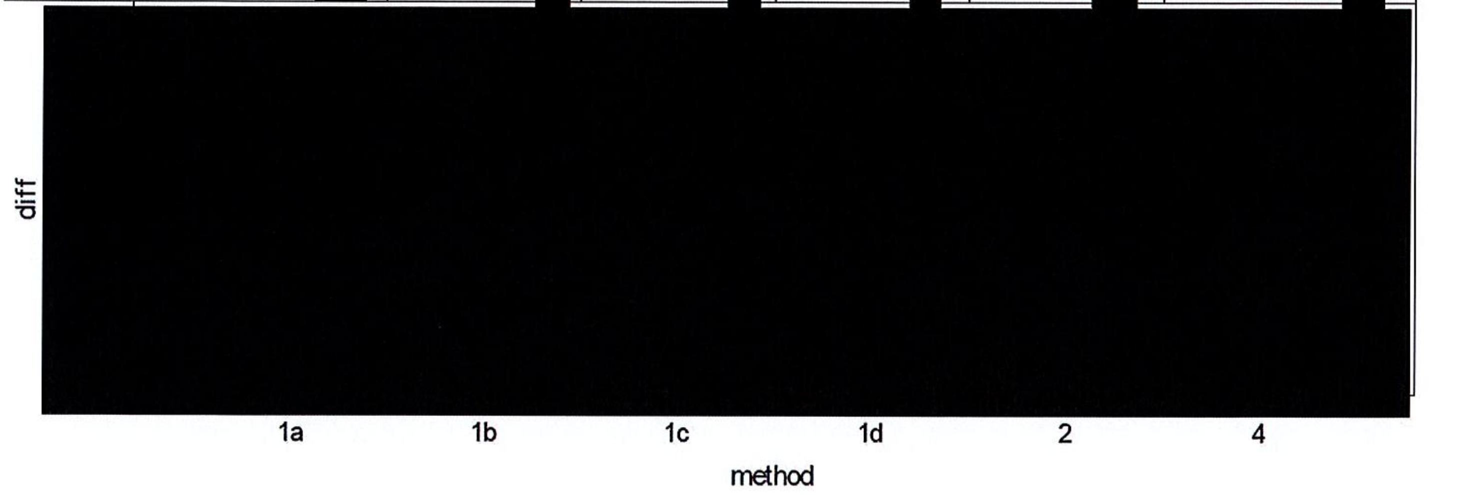

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=702

Overall Statistics					
Min	[redacted]	Mean	-7.[redacted]	Max	[redacted]
Pooled Std Dev	120.[redacted]				

Statistics by GQtype						
N	750	750	750	750	750	750
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-2.[redacted]	-12.[redacted]	-9.[redacted]	-18.[redacted]	-1.[redacted]	1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	1100	2300	5700	2400	1800	1600
Std Dev	64.[redacted]	112.[redacted]	205.[redacted]	118.[redacted]	78.[redacted]	92.[redacted]


[redacted]

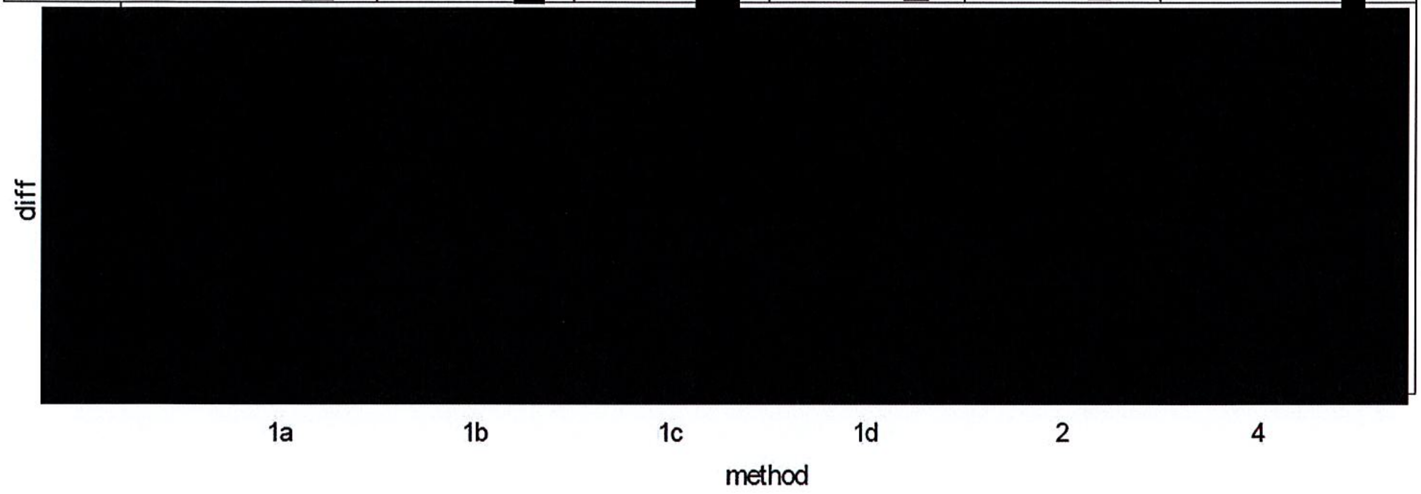
DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=704

Overall Statistics					
Min	[redacted]	Mean	-5.[redacted]	Max	[redacted]
Pooled Std Dev	69.[redacted]				

Statistics by GQtype

N	30	30	30	30	30	30
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	3.[redacted]	-20.[redacted]	-0.[redacted]	-20.[redacted]	-0.[redacted]	6.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	100	750	100	750	90	100
Std Dev	21[redacted]	116.[redacted]	17.[redacted]	116.[redacted]	17.[redacted]	23.[redacted]

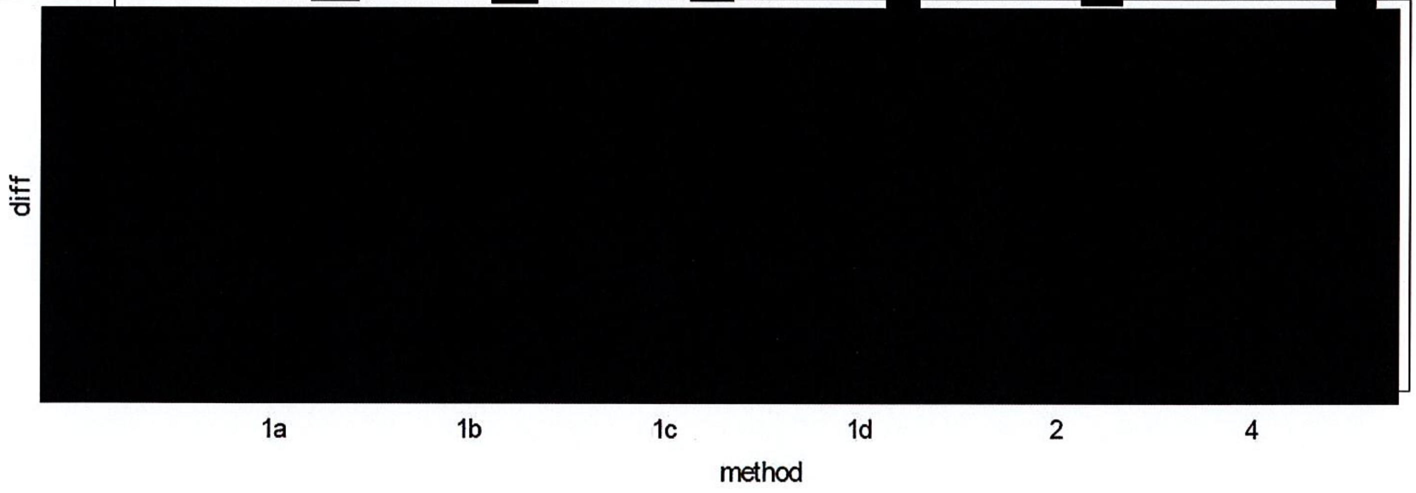

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=706

Overall Statistics					
Min	[redacted]	Mean	-1.[redacted]	Max	[redacted]
Pooled Std Dev	24.[redacted]				

Statistics by GQtype						
N	200	200	200	200	200	200
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-4.[redacted]	-4.[redacted]	-0.[redacted]	-0.[redacted]	1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	150	200	600	150	200	250
Std Dev	18.[redacted]	22.[redacted]	42.[redacted]	16.[redacted]	18.[redacted]	19.[redacted]



[redacted]

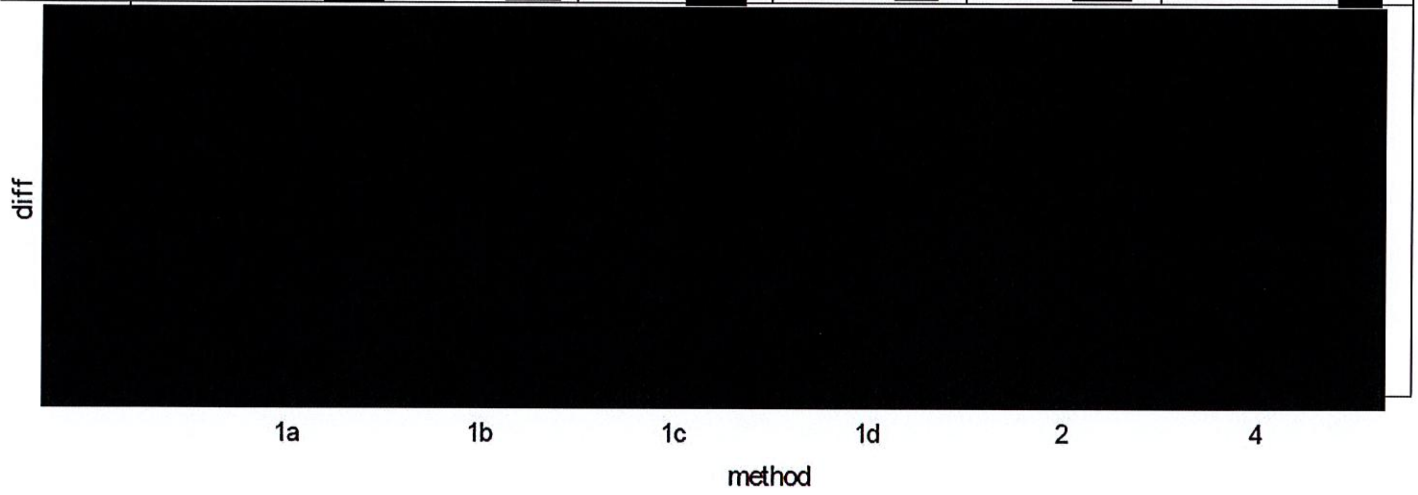
DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=801

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	14.[redacted]				

Statistics by GQtype						
N	21000	21000	21000	21000	21000	21000
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	-0.[redacted]	1.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	300	2200	650	2100	600	500
Std Dev	5.[redacted]	19.[redacted]	9.[redacted]	19.[redacted]	9.[redacted]	14.[redacted]

[redacted]

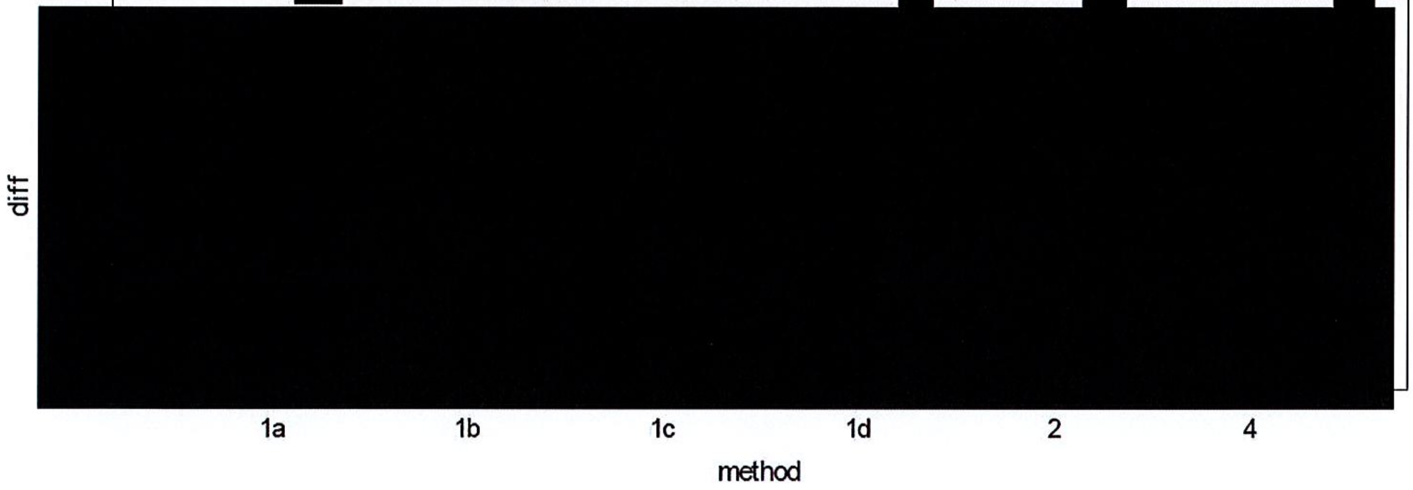


DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=802

Overall Statistics					
Min	[redacted]	Mean	-0.[redacted]	Max	[redacted]
Pooled Std Dev	20.[redacted]				

Statistics by GQtype						
N	3600	3600	3600	3600	3600	3600
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-0[redacted]	-0.[redacted]	-0[redacted]	-0.[redacted]	-0.[redacted]	2.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	450	850	650	850	650	550
Std Dev	13.[redacted]	17.[redacted]	20.[redacted]	17.[redacted]	15.[redacted]	31.[redacted]


[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=901

Overall Statistics					
Min	[redacted]	Mean	-9.[redacted]	Max	[redacted]
Pooled Std Dev	757.[redacted]				

Statistics by GQtype

N	2300	2300	2300	2300	2300	2300
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-1.[redacted]	-27.[redacted]	1.[redacted]	-30.[redacted]	-0.[redacted]	3.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	500	58500	650	66500	500	600
Std Dev	20.[redacted]	1224	27.[redacted]	1392	20.[redacted]	34.[redacted]

diff

[redacted]

1a 1b 1c 1d 2 4

method

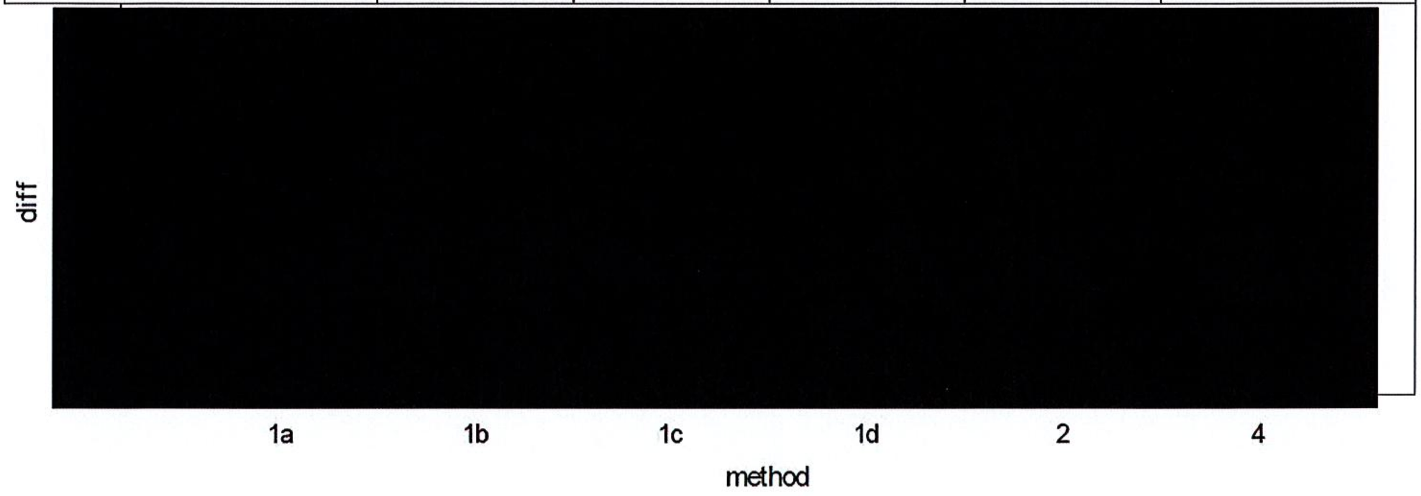
DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=903

Overall Statistics					
Min	[redacted]	Mean	-2.[redacted]	Max	[redacted]
Pooled Std Dev	28.[redacted]				

Statistics by GQtype						
N	<15	<15	<15	<15	<15	<15
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	7	-2	-3	0	0	-18
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	40	<15	<15	<15	<15	150
Std Dev	20	3	5	2	2	67

[redacted]



DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

GQTYPCUR=999

Overall Statistics					
Min	[redacted]	Mean	-5.[redacted]	Max	[redacted]
Pooled Std Dev	39.[redacted]				

Statistics by GQtype						
N	20	20	20	20	20	20
Min	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q1	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Q2	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Mean	-16[redacted]	-7.[redacted]	-17.[redacted]	-1.[redacted]	-2.[redacted]	12.[redacted]
Q3	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Max	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Range	200	90	250	30	200	100
Std Dev	49.[redacted]	20[redacted]	63.[redacted]	8.[redacted]	38[redacted]	32.[redacted]

diff

[redacted]

1a 1b 1c 1d 2 4

method

DRB Approval Number: CBDRB-FY21-DSEP-002
Statistics have been rounded according to Census Bureau disclosure standards

Table 6: Population of Group Quarters by Group Quarters Category

Types of GQs		Group Quarters		Population		
		Count*	Percent of Total+	Count*		Percent of Total+
Total...............................	167,000	[redacted]	100.00	[redacted]	8,025,000	100.00
College/University Student Housing......................	28,000	[redacted]	16.74	[redacted]	2,524,000	31.45
Correctional Facilities for Adults	12,500	[redacted]	7.38	[redacted]	2,277,000	28.37
Group Homes Intended for Adults......	40,500	[redacted]	24.19	[redacted]	307,000	3.83
Hospitals** and In Patient Hospices	1,900	[redacted]	1.15	[redacted]	71,000	0.88
Juvenile Facilities..................	9,200	[redacted]	5.49	[redacted]	153,000	1.90
Living Quarters for Victims of Natural Disasters	N<15	[redacted]	0.00	[redacted]	30	0.00
Military Quarters................	2,900	[redacted]	1.75	[redacted]	289,000	3.60
Military/Maritime Vessels	450	[redacted]	0.26	[redacted]	52,000	0.65
Nursing and Skilled Nursing Facilities...........................	22,000	[redacted]	13.04	[redacted]	1,508,000	18.79
Religious Group Quarters and Domestic Violence Shelters	10,500	[redacted]	6.32	[redacted]	101,000	1.26
Residential Schools for People with Disabilities....................	350	[redacted]	0.19	[redacted]	9,700	0.12
Residential Treatment Centers for Adults.........................	8,200	[redacted]	4.91	[redacted]	142,000	1.77
Shelters and Service-based locations.......	18,500	[redacted]	11.11	[redacted]	423,000	5.27
Workers' Group Living Quarters and Job Corp Centers............	12,500	[redacted]	7.47	[redacted]	169,000	2.11

*Counts and percentages are unweighted.

+Percentages may not sum to 100 due to rounding.

**Hospitals include GQs that were mental or psychiatric hospitals, the mental or psychiatric unit or floor for long term care at a regular hospital or hospitals that accept patients with no disposition.

Source: 2010 Census Edited File (CEF)

County Distribution of 2020 Census / 2020 ACS - GQ Person Ratios Before Imputation

[redacted]

County Distribution of 2020 Census / 2010 ACS - GQ Person Ratios After Imputation

[redacted]

County Distribution of 2020 Census / 2020 ACS - GQ Person Ratios Before Imputation

[redacted]

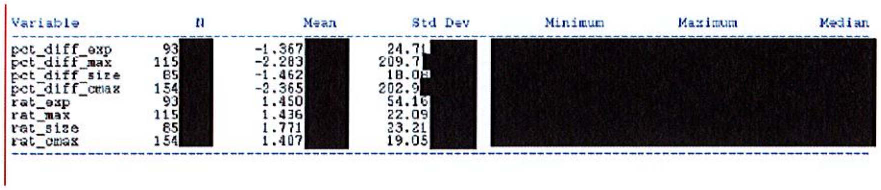
DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

County Distribution of 2020 Census / 2010 ACS - GQ Person Ratios After Imputation

[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

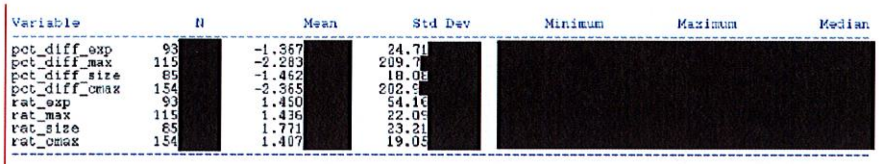
Military	0.7533	0.2347	1.047	0 3326
Shelters	0.6875	0.5768	0.6762	0.6070
Group Homes	0.8866	0.5374	1.048	0 5307
Other	0.8283	0.4205	1.084	0.4002
All GQs	0.8211	0.5377	0.9851	0 5478



```
Variable              N              Mean          Std Dev          Minimum          Maximum          Median
-----------------------------------------------------------------------------------------------------------
pct_diff_exp         93 [redacted] -1.367 [redacted] 24.7 [redacted] [redacted]
pct_diff_max        115            -2.283            209.7
pct_diff_size        85            -1.462            18.0
pct_diff_cmax       154            -2.365            202.9
rat_exp              93             1.450            54.16
rat_max             115             1.436            22.09
rat_size             85             1.771            23.21
rat_cmax            154             1.407            19.05
-----------------------------------------------------------------------------------------------------------
```

Commented [JEZ(F10]: I can format this tomorrow. Key: pct_diff = (GP – Imputed value)/GP for resolved cases. Rat = GP/AUX. Exp = GQAC Expected Count
Max = GQAC Max Number of People
Size = Current GQ Size
CMax = Max Number of People

Military	0.7533	0.2347	1.047	0 3326
Shelters	0.6875	0.5768	0.6762	0.6070
Group Homes	0.8866	0.5374	1.048	0 5307
Other	0.8283	0.4205	1.084	0.4002
All GQs	0.8211	0.5377	0.9851	0 5478



```
Variable              N            Mean          Std Dev          Minimum          Maximum           Median
------------------------------------------------------------------------------------------------------------
pct_diff_exp         93[redacted]   -1.367[redacted]   24.71[redacted]   [redacted]
pct_diff_max        115[redacted]   -2.283[redacted]  209.7[redacted]    [redacted]
pct_diff_size        85[redacted]   -1.462[redacted]   18.0[redacted]    [redacted]
pct_diff_cmax       154[redacted]   -2.365[redacted]  202.[redacted]     [redacted]
rat_exp              93[redacted]    1.450[redacted]   54.1[redacted]    [redacted]
rat_max             115[redacted]    1.436[redacted]   22.0[redacted]    [redacted]
rat_size             85[redacted]    1.771[redacted]   23.21[redacted]   [redacted]
rat_cmax            154[redacted]    1.407[redacted]   19.05[redacted]   [redacted]
------------------------------------------------------------------------------------------------------------
```

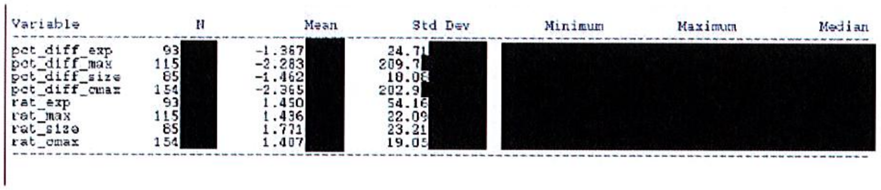
Commented [JEZ(F6]: I can format this tomorrow. Key: pct_diff = (GP – Imputed value)/GP for resolved cases. Rat = GP/AUX. Exp = GQAC Expected Count
Max = GQAC Max Number of People
Size = Current GQ Size
CMax = Max Number of People

GQAC Max Number of People	Resolved	Unresolved	Total
Populated	116,000	15,000	131,000
Not Populated	68,000	28,500	96,000
Total	184,000	43,000	227,000

Current GQ Size	Resolved	Unresolved	Total
Populated	154,000	29,500	183,000
Not Populated	29,500	14,000	43,000
Total	184,000	43,000	227,000

Max Number of People	Resolved	Unresolved	Total
Populated	85,000	14,500	99,500
Not Populated	98,500	28,500	127,000
Total	184,000	43,000	227,000

GQ Type	Ratio of GQAC Expected Count to Good People Count	Ratio of GQAC Max Number of People to Good People Count	Ratio of Current GQ Size to Good People Count	Ratio of Max Number of People to Good People Count
Correctional Facilities	0.7239	0.4350	0.9223	0.4468
Juvenile Facilities	0.6977	0.3075	0.8621	0 3284
Nursing Facilities	0.8875	0.6815	0.9705	0.6810
Hospitals	0.7779	0.6424	1.017	0.6441
College Housing	0.8203	0.6147	1.071	0.6114
Military	0.7533	0.2347	1.047	0 3326
Shelters	0.6875	0.5768	0.6762	0.6070
Group Homes	0.8866	0.5374	1.048	0 5307
Other	0.8283	0.4205	1.084	0.4002
All GQs	0.8211	0.5377	0.9851	0 5478



```
Variable              N              Mean          Std Dev          Minimum          Maximum          Median
-------------------------------------------------------------------------------------------------------------
pct_diff_exp         93 [redacted]   -1.367 [redacted]  24.71 [redacted]  [redacted]
pct_diff_max        115              -2.283            209.7
pct_diff_size        85              -1.462             10.0
pct_diff_cmax       154              -2.365            202.9
rat_exp              93               1.450             54.16
rat_max             115               1.436             22.09
rat_size             85               1.771             23.21
rat_cmax            154               1.407             19.05
-------------------------------------------------------------------------------------------------------------
```

Commented [JEZ(F7]: I can format this tomorrow. Key: pct_diff = (GP – Imputed value)/GP for resolved cases. Rat = GP/AUX. Exp = GQAC Expected Count
Max = GQAC Max Number of People
Size = Current GQ Size
CMax = Max Number of People

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

from the college-level sum of GQ population counts for at least three reasons: (1) **reference year**—our latest IPEDS data is for reference year 2019; (2) **"capacity utilization"**—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus, while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day; (3) **scope**—IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Adjusting the IPEDS facility-level Room Capacity

To adjust the IPEDS room capacity for reference year differences, we use the GQAC Max Number of People. We first select colleges for which we have a positive GQAC Max Number of People for every GQ at the facility. Since the IPEDS data does not include off-campus housing, we further subset on facilities that have no Greek letter GQs (fraternity or sorority houses). Finally, to maximize the chances that we are comparing apples to apples, we also subset to facilities for which the match quality is very high (match score > 90%). Within this subset, we calculate the average ratio of the facility-level sum of GQAC Max Number of People over the room capacity from IPEDS:

$$Average\ Ratio_S = \sum_{i \in S} \frac{\sum_{facility\_i} GQAC\ Max\ Number\ of\ People}{IPEDS\ Room\ Capacity\ at\ facility\ i}$$

where *S* is the set of facilities with no Greek GQs only positive values for GQAC Max Number of People.

Reassuringly, within this set of facilities, the median ratio is [redacted], the mode is [redacted], the 25th percentile is [redacted], and the 75th percentile is [redacted].

After adjusting the IPEDS college-level room capacity, we will similarly adjust for GQ "capacity utilization" at the college-level, using the mean ratio of 2020 Census Day GQ population over GQAC Max Number of People for all GQs for which both 2020 Census Day GQ population over GQAC Max Number of People. If time and sample sizes permit, we will also calculate this average ratio for college size classes. If the mean ratios differ significantly by college size class we will use separate capacity utilization adjustment for each college size class

After adjusting the college-level total room capacity to account reference year for capacity utilization, we will calculate the following college-level residual for each college C:

$$Residual_C = Adjusted\ IPEDS\ Room\ Capacity_C - \sum_{C} Reported\ GQ\ Pop\ Count - \sum_{C*} GQAC\ Expected\ Count$$

where the first summation is over all GQs at college C with a good person count, and the second summation is over all GQs at college C *without* a good person count but with positive GQAC Expected Count.

Finally, we will adjust the room capacity for GQ population in off-campus Greek housing (which is not included in the IPEDS room capacity). About 51% of colleges in the GQ data have no Greek letter GQs.

Adjusted Residual from Facility-level Total for College Housing

Formatted: Normal

If the GQ advance contact expected count is not populated, we will implement the following facility-level residual method. This method can only be used for GQs with GQTYPCUR=501 (colleges and universities)~~.~~ (For the rest of this section we use "college" "university" or "facility" to mean the same thing.

For ~~universities and colleges~~501s**, we have the 2019** ~~facility~~college**-level total room capacity (number of persons that could live in the GQ) from the IPEDS. This has been matched at the** ~~facility~~ college **level to the GQ data.** The main advantage of this variable is that it is available for over 99% of the 501 type GQs. The IPEDS room capacity may differ from the college-level sum of GQ population counts for at least three reasons: (1) **reference year**—our latest IPEDS data is for reference year 2019; (2) "**capacity utilization**"—the IPEDS data is for the maximum number of persons that *could* live in all the GQs on campus while the Census Day GQ population count should only include persons who would normally be in the GQ on Census Day; (3) **scope**—IPEDS includes only on-campus housing, while the GQ data includes off-campus fraternity and sorority houses. We adjust the college-level room capacities for each of these factors.

Commented [TKW(F14]: Joe Staudt can provide a description of the matching algorithm, and the quality of the matches (which is very high for a high percentage of the cases).

Formatted: Font: Bold

Formatted: Font: Bold

Formatted: Font: Italic

Formatted: Font: Bold

Adjusting the IPEDS facility-level Room Capacity

To adjust the IPEDS room capacity for reference year differences we use the GQAC Max Number of People. We first select colleges for which we have a positive GQAC Max Number of People for every GQ at the facility. Since the IPEDS data does not include off-campus housing, we further subset on facilities that have no Greek letter GQs (fraternity or sorority houses). Finally, to maximize the chances that we are comparing apples to apples we also subset to facilities for which the match quality is very high (match score > 90%). Within this subset we calculate the average ratio of the facility-level sum of GQAC Max Number of People over the room capacity from IPEDS:

Formatted: Font: 11 pt

Formatted: Font: 11 pt

Formatted: Font: 11 pt

$$Average\ Ratio_S = \sum_{i \in S} \frac{\sum_{facility\_i} GQAC\ Max\ Number\ of\ People}{IPEDS\ Room\ Capacity\ at\ facility\ i}$$

where *S* is the set of facilities with no Greek GQs only positive values for GQAC Max Number of People.

Reassuringly, within this set of facilities, the median ratio is [redacted], the mode is [redacted], the 25th percentile is [redacted] and the 75th percentile is [redacted].

Formatted: Font: 11 pt

Formatted: Font: 11 pt

After adjusting the IPEDS college-level room capacity we will similarly adjust for GQ "capacity utilization" at the college-level using the mean ratio of 2020 Census Day GQ population over GQAC Max Number of People for all GQs for which both 2020 Census Day GQ population over GQAC Max Number of People. If time and sample sizes permit, we will also calculate this average ratio for college size classes. If the mean ratios differ significantly by college size class we will use separate capacity utilization adjustment for each college size class

Formatted: Font: 11 pt

Formatted: Font: 11 pt

After adjusting the college-level total room capacity to account reference year for capacity utilization we will calculate the following college-level residual for each college C:

shown in Table 9. Tables 12-14 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 9: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0.3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1.017	0.6385
College Housing	0.7818	0.5492	0.9444	0.5535
Military	0.7317	0.2290	0.9492	0.2914
Shelters	0.6261	0.5325	0.6180	0.5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0.3597
All GQs	0.7878	0.5057	0.9217	0.5153

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for GQs for colleges and universities (GQTYPCUR=501).

Adjusting the IPEDS facility-level Room Capacity

To adjust the IPEDS room capacity for reference year differences, we use the GQAC Max Number of People. We first select colleges for which we have a positive GQAC Max Number of People for every GQ at the facility. Since the IPEDS data does not include off-campus housing, we further subset on facilities that have no Greek letter GQs (fraternity or sorority houses). Finally, to maximize the chances that we are comparing apples to apples, we also subset to facilities for which the match quality is very high (match score > 90%). Within this subset, we calculate the average ratio of the facility-level sum of GQAC Max Number of People over the room capacity from IPEDS:

$$Average\ Ratio_S = \sum_{i \in S} \frac{\sum_{college_i} GQAC\ Max\ Number\ of\ People}{IPEDS\ Room\ Capacity\ at\ college\ i}$$

where *S* is the set of colleges with no Greek GQs only positive values for GQAC Max Number of People.

Reassuringly, within this set of colleges, the median ratio is [redacted], the mode is [redacted], the 25th percentile is [redacted], and the 75th percentile is [redacted].

After adjusting the IPEDS college-level room capacity, we will similarly adjust for GQ "capacity utilization" at the college-level, using the mean ratio of 2020 Census Day GQ population over GQAC Max Number of People ~~for all~~among ~~GQs~~ colleges for which ~~both~~all 2020 Census Day GQ populations ~~over~~ and GQAC Max Number of People are non-missing and positive. If time and sample sizes permit, we will also calculate this average ratio for college size classes. If the mean ratios differ significantly by college size class we will use separate capacity utilization adjustment for each college size class

Commented [TLK(F14]: I think a word is missing from this sentence.

Commented [TKW(F15]: I have added the missing words now.

shown in Table 9. Table 12Table 14 in the Appendix show counts of populated records for which these ratio methods could be used.

Table 9: Factors to convert Auxiliary Variables to GQ Population

GQ Type	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Correctional Facilities	0.7181	0.4332	0.9174	0.4450
Juvenile Facilities	0.6734	0.2974	0.8369	0 3175
Nursing Facilities	0.8617	0.6603	0.9408	0.6591
Hospitals	0.7709	0.6391	1.017	0.6385
College Housing	0.7818	0.5492	0.9444	0 5535
Military	0.7317	0.2290	0.9492	0 2914
Shelters	0.6261	0.5325	0.6180	0 5689
Group Homes	0.8299	0.5009	0.9679	0.4996
Other	0.7384	0.3783	0.9276	0 3597
All GQs	0.7878	0.5057	0.9217	0 5153

Adjusted Residual from Facility-level Total for College Housing

A second imputation method under consideration is the Adjusted Residual from Facility-level Totals for College Housing. This method can only be used for GQs for colleges and universities (GQTYPCUR=501).

Adjusting the IPEDS facility-level Room Capacity

To adjust the IPEDS room capacity for reference year differences, we use the GQAC Max Number of People. We first select colleges for which we have a positive GQAC Max Number of People for every GQ at the facility. Since the IPEDS data does not include off-campus housing, we further subset on facilities that have no Greek letter GQs (fraternity or sorority houses). Finally, to maximize the chances that we are comparing apples to apples, we also subset to facilities for which the match quality is very high (match score > 90%). Within this subset, we calculate the average ratio of the facility-level sum of GQAC Max Number of People over the room capacity from IPEDS:

$$Average\ Ratio_S = \sum_{i\in S} \frac{\sum_{college_i} GQAC\ Max\ Number\ of\ People}{IPEDS\ Room\ Capacity\ at\ college\ i}$$

where *S* is the set of colleges with no Greek GQs only positive values for GQAC Max Number of People.

Reassuringly, within this set of colleges, the median ratio is [redacted], the mode is [redacted], the 25th percentile is [redacted], and the 75th percentile is [redacted].

After adjusting the IPEDS college-level room capacity, we will similarly adjust for GQ "capacity utilization" at the college-level, using the mean ratio of 2020 Census Day GQ population over GQAC Max Number of People for all GQs for which both 2020 Census Day GQ population over GQAC Max Number of People. If time and sample sizes permit, we will also calculate this average ratio for college size classes. If the mean ratios differ significantly by college size class we will use separate capacity utilization adjustment for each college size class

Commented [TLK(F14]: I think a word is missing from this sentence.

HB-Edits based on Expected Count:

GQ Type	Count not Flagged	Impute	Suppress	Review	No Expected Count or no GP	Total
Correctional Facilities	3,300	50	N<15	20	12,500	16,000
Juvenile Facilities	3,700	50	20	80	4,200	8,000
Nursing Facilities	19,000	250	20	50	9,200	28,500
Hospitals	1,200	30	20	20	1,500	2,800
College Housing	17,500	800	300	150	17,500	36,000
Military	950	20	N<15	30	4,000	5,000
Shelters	3,600	40	40	60	29,000	33,000
Group Homes	33,000	400	70	100	38,500	72,000
Other	8,900	150	30	40	17,000	26,000
All GQs	90,500	1,700	500	550	133,000	227,000

Commented [JEZ(F1]: Note, this is sum of 90,000 + 8,600 + 34,500 in Table 8.

[redacted]

Commented [JEZ(F2]: Example plot for Hospitals...(because it's easiest to see)

X = log(gp), Y = log(expected count). Ignore 'M' in legend.

I = Impute
S = Suppress
R = Review

GQ Type	Ratio of Good People Count to GQAC Expected Count [ORIG]	Ratio of Good People Count to GQAC Expected Count [EDIT]
Correctional Facilities	0.7181	0.7562
Juvenile Facilities	0.6734	0.7318
Nursing Facilities	0.8617	0.8712
Hospitals	0.7709	0.8090
College Housing	0.7818	0.9045
Military	0.7317	0.7579
Shelters	0.6261	0.6330
Group Homes	0.8299	0.8801
Other	0.7384	0.8719
All GQs	0.7878	0.8498

[redacted]

Commented [JEZ(F3]: Histogram of ratios of GP/Expected count. Bounded by [redacted]

Group Quarters Imputation Methodology

GQ Greek Indicator	GQ Count
Greek Indicator = 1	4,100
Greek Indicator = 0*	35,000
All 501s	39,500

* includes GQTYPE = 501s where GQ name is missing.

GQ Greek Indicator	Ratio of Good People Count to GQAC Expected Count	Ratio of Good People Count to GQAC Max Number of People	Ratio of Good People Count to Current GQ Size	Ratio of Good People Count to Max Number of People
Greek Indicator = 1	0.8327	0.5443	0.8807	0.2618
Greek Indicator = 0*	0.7800	0.7216	0.9472	0.5749
All 501s	0.7818	0.5492	0.9444	0.5535

* includes GQTYPE = 501s where GQ name is missing.

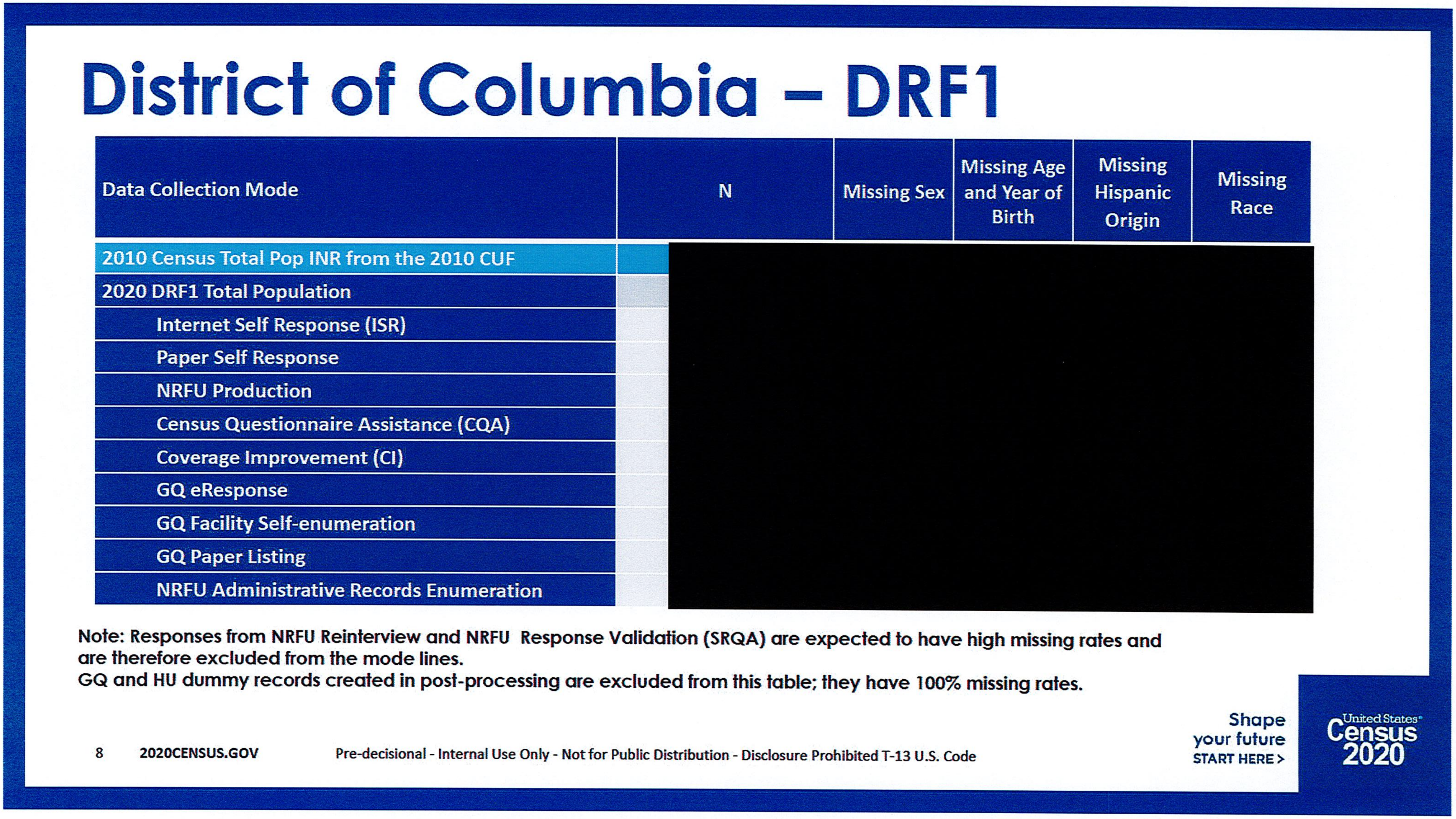
GQ Type	Median Good People Count
Correctional Facilities	[redacted]
Juvenile Facilities	[redacted]
Nursing Facilities	[redacted]
Hospitals	[redacted]
College Housing	[redacted]
Military	[redacted]
Shelters	[redacted]
Group Homes	[redacted]
Other	[redacted]
All GQs	[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

Person Information

Analysis of 39 Schools

- 139,857 person records
    - [redacted]
- Seeing fairly complete reporting of first and last name
- 7 schools did not provide date of birth
- 7 schools provided sex
- 9 schools provided information in race field
- 7 schools provided information in Hispanic field



District of Columbia – DRF1

Data Collection Mode	N	Missing Sex	Missing Age and Year of Birth	Missing Hispanic Origin	Missing Race
2010 Census Total Pop INR from the 2010 CUF	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
2020 DRF1 Total Population	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Internet Self Response (ISR)	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Paper Self Response	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
NRFU Production	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Census Questionnaire Assistance (CQA)	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Coverage Improvement (CI)	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
GQ eResponse	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
GQ Facility Self-enumeration	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
GQ Paper Listing	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
NRFU Administrative Records Enumeration	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Note: Responses from NRFU Reinterview and NRFU Response Validation (SRQA) are expected to have high missing rates and are therefore excluded from the mode lines.
GQ and HU dummy records created in post-processing are excluded from this table; they have 100% missing rates.

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

South Carolina - Hispanic Origin and Race Reasonableness

[redacted]

Not Hispanic Hispanic No Response

🟩 2010 CUF 🟩 2019 Estimates 🟦 2020 DRF1

White alone Black alone AIAN alone Asian alone NHPI alone SOR alone Two or More Races No Response

🟩 2010 CUF 🟩 2019 ACS 🟦 2020 DRF1

Note: Hispanic origin and race groups determined by checkbox responses only.

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

POP's Approach for GQ Type 501 (College/University Housing)

[redacted]

Summarizing the Map…

There are [redacted] census tracts in [redacted] states that meet these criteria.

- [redacted] of these tracts have a percentage decline of 90% or more.
- tracts have a 100% decline in population from the benchmark data.

- [redacted] tracts have a population decline of 1,000 or more.
- tracts have a population decline of 2,000 or more.

It is not clear that all of these colleges identified are actually problematic. Additional research into these GQs must be done in order to verify. There may be valid reasons why some of these losses occurred. For examples, colleges or individual dorms may have closed in years prior to the 2020 Census.

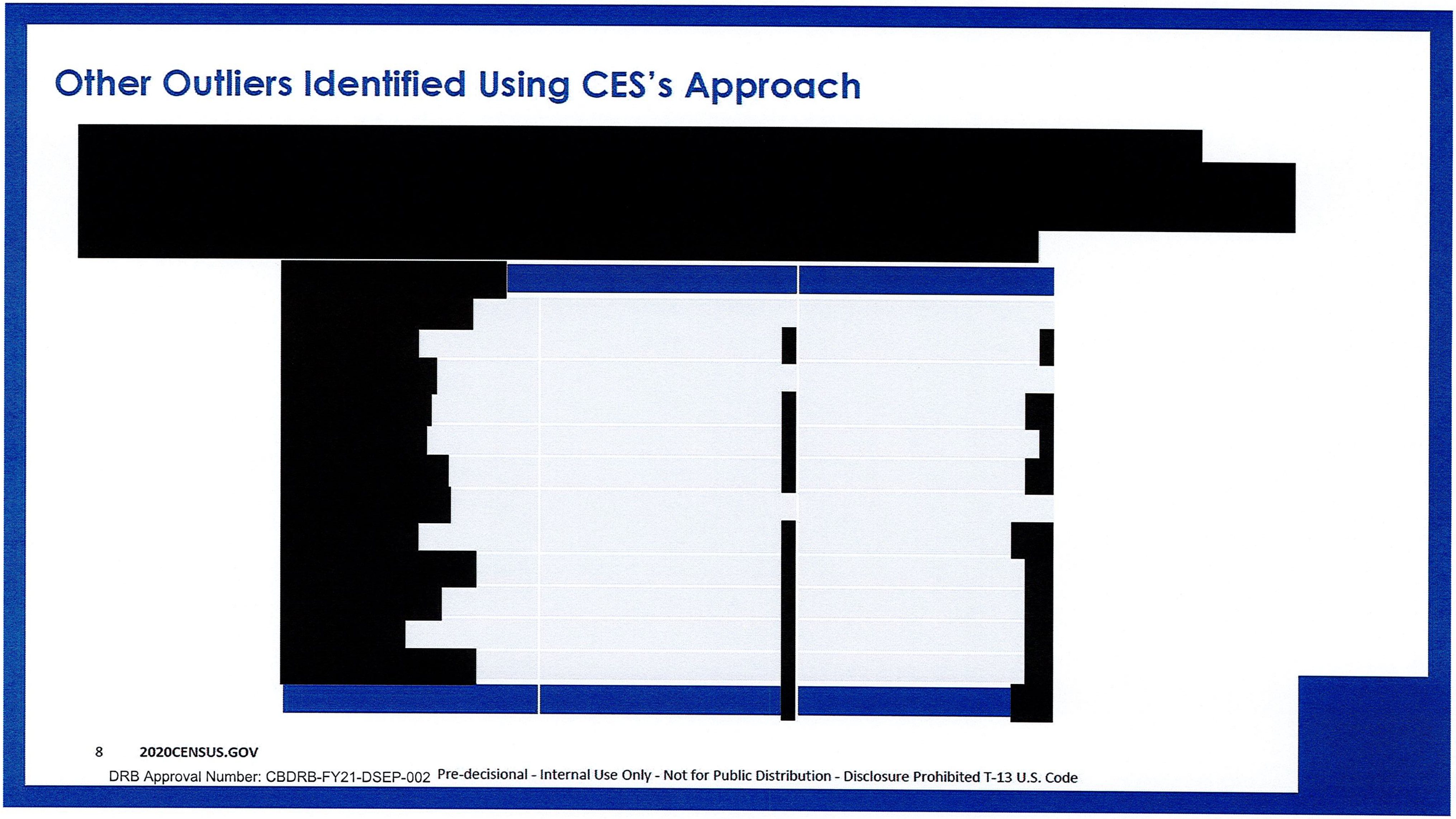
[redacted]

Census Tracts with 100% Decline from 2013-2017 ACS

State	County	Tract GEOID	Facility Name
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
		[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]	[redacted]
		[redacted]	[redacted]

DRB Approval Number: CBDRB-FY21-DSEP-002

Pre-decisional - Internal Use Only - Not for Public Distribution - Disclosure Prohibited T-13 U.S. Code



Other Outliers Identified Using CES's Approach

[redacted]

[redacted]		
[redacted]		
[redacted]	[redacted]	[redacted]
[redacted]		
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]		
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
	[redacted]	[redacted]

POP's Approach – Prisons – [redacted] Census Tracts

Group Quarters Anomalies: Type 100s (Adult Correctional Facilities)

[redacted]

Pre-decisional - Internal Use Only - Not for Public Distribution - Disclosure Prohibited T-13 U.S. Code

POP's Approach – Military Quarters – [redacted] Census Tracts

Group Quarters Anomalies: Type 600s (Military Quarters)

[redacted]

POP's Approach for GQ Type 501 90%+

Group Quarters Anomalies: Type 501 (College/University Housing)

[redacted]

POP's Approach for GQ Type 501 – 100%+

Group Quarters Anomalies: Type 501 (College/University Housing)

[redacted]

County Distribution of 2020 Census / 2020 ACS - GQ Person Ratios Before Imputation

[redacted]

County Distribution of 2020 Census / 2010 ACS - GQ Person Ratios After Imputation

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

[redacted]

[redacted]

[redacted]

[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

[redacted]

[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]

[redacted]

[redacted]

These are the tracts with the GQ Prison 100s types.
It looks at the 25 largest BCU geography tracts where 99 percent of the MAFID matches to another GQ.

One thing that stands out is that several of these tracts have only one GQ MAFID.

- In looking at [redacted], this MAFID has almost everyone in another GQ in a different tract.

Table x: BCU Geography Tracts with Largest number of Persons in MAFIDs with 99+ percent match rate

Obs	BCUSTATEFP	BCUCOUNTYFP	BCUTRACTCE	_FREQ_	sum_scount
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	3,100
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	950
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	900
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	850
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	750
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	650
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	350
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	350
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	350
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	14,000

This is the 25 largest tracts when doing for Nursing Home 300s

These are MAFIDs where over 99 percent are found in another group quarters

Similar results as prisons. These tracts have only one MAFID. When this is happening that the people are found in a MAFID that is in a different tract.

Table 4: 25 Tracts with Largest Number of Nursing Home People Found in a Group Quarters

Obs	BCUSTATEFP	BCUCOUNTYFP	BCUTRACTCE	_TYPE_	_FREQ_	sum_scount
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	100
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	100
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	100
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	4,800

Tuesday, December 8, 2020

This is College Dorms GQ Types 500s

Table 4: Largest 25 Tracts for College Dorms GQ 500s where over 99 percent of the people in the MAFD match to another GQ

These could possibly be candidates for doing what was just done for the previous patch.

Obs	BCUSTATEFP	BCUCOUNTYFP	BCUTRACTCE	_TYPE_	_FREQ_	sum_scount
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	350
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	150
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	13,000

DRB Approval Number: CBDRB-FY21-DSEP-002 Statistics have been rounded according to Census Bureau disclosure standards

These are the tracts with the GQ Prison 100s types.
It looks at the 25 largest BCU geography tracts where 99 percent of the MAFID matches to another GQ.

One thing that stands out is that several of these tracts have only one GQ MAFID.
The 12/09/20 correction is showing that the top 25 now generally have 2+ MAFIDs in the tract.

Sum_count is the number of people in the MAFIDs in the tact that have 99+ percent match rate

Table x: BCU Geography Tracts with Largest number of Persons in MAFIDs with 99+ percent match rate

Obs	BCUSTATEFP	BCUCOUNTYFP	BCUTRACTCE	_TYPE_	_FREQ_	sum_scount
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	5,600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	2,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,900
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,800
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	950
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	900
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	850
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	750
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	750
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	650
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	650
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	550
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	29,000

This is the 25 largest tracts when doing for Nursing Home 300s

These are MAFIDs where over 99 percent are found in another group quarters

Similar results as prisons. The 12/09/20 rerun is finding a second MAFID in the tract.

Table 4: 25 Tracts with Largest Number of Nursing Home People Found in a Group Quarters

Obs	BCUSTATEFP	BCUCOUNTYFP	BCUTRACTCE	_TYPE_	_FREQ_	sum_scount
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	650
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	550
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	550
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	250
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	200
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	8,300

This is College Dorms GQ Types 500s

Table 4: Largest 25 Tracts for College Dorms GQ 500s where over 99 percent of the people in the MAFD match to another GQ

These could possibly be candidates for doing what was just done for the previous patch. Additonal MAFIDs and population were identified by 12/09/20 rerun

Obs	BCUSTATEFP	BCUCOUNTYFP	BCUTRACTCE	_TYPE_	_FREQ_	sum_scount
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	4,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	3,400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	2,700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	2,400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	2,100
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,800
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,300
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,100
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	1,000
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	900
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	750
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	700
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	600
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	500
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	450
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	400
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	350
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	32,000

10 Counties with Highest % HHs 18-29 & Enrolled in Post-Secondary Education

% HHs Age 18-29 College	County	State	University
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]

Pre-Decisional: Internal Use Only

DRB Approval Number: CBDRB-FY21-DSEP-002. Statistics have been rounded according to Census Bureau disclosure standards.

UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA

FAIR LINES AMERICA FOUNDATION, INC., Plaintiff, v. UNITED STATES DEPARTMENT OF COMMERCE and UNITED STATES BUREAU OF THE CENSUS, Defendants.	Case No. 1:21-cv-1361-ABJ

[PROPOSED] ORDER GRANTING PRELIMINARY INJUNCTION

Based upon the pleadings, motions, and evidence received by the Court, the Court hereby GRANTS the motion filed by Plaintiff seeking a preliminary injunction and ORDERS as follows:

1. Defendants are hereby enjoined from failing to comply with Plaintiff's March 31, 2021 FOIA request, and are ordered to produce all responsive non-exempt records and data improperly withheld from the May 25 production within 10 days of the date of the Court's Order, [or before August 15, 2021, whichever is earlier], including tabulations and statistical materials that do not disclose the information reported by or on behalf of, any particular respondent, which includes intermediate work product and data that was not furnished by any particular establishment or individual while excluding personally identifiable information.

2. Defendants are hereby enjoined from failing to timely respond to Plaintiff's Request pursuant to 5 U.S.C. § 552(a)(6)(A) regarding Defendants' identified potentially responsive emails, and are ordered to identify and produce all non-exempt responsive email

records to Plaintiff as soon as practicable.

3. Defendants are also ordered to produce a *Vaughn* Index specifically describing in detail each record and portion thereof withheld as exempt within the same timeframe.

This Order shall remain in effect through the remainder of these proceedings until such time as the Court enters a subsequent Order dissolving this preliminary injunction and/or awarding permanent relief.

Date: ________________

____

United States District Judge

CERTIFICATE OF SERVICE

I do hereby certify that pursuant to LCvR 7(k), on this 19th day of July 2021 the foregoing Proposed Order was filed electronically with the Clerk of Court using the CM/ECF system. The system instantaneously generated a Notice of Electronic Filing which served all counsel of record.

/s/ Jason Torchinsky
Jason Torchinsky (D.C. Bar No. 976033)
jtorchinsky@hvjt.law
Jonathan P. Lienhard (D.C. Bar No. 474253)
jlienhard@hvjt.law
Kenneth C. Daines (D.C. Bar No. 1600753)*
*Pro hac vice motion pending
kdaines@hvjt.law
HOLTZMAN VOGEL JOSEFIAK TORCHINSKY PLLC
15405 John Marshall Highway
Haymarket, VA 20169
Phone: (540) 341-8808
Counsel for Plaintiff