Harvard Data Science Review • Issue 2.1, Winter 2020

Distrust in the Governments Brings Risk to the Census

Howard Hogan¹

¹United States Census Bureau, Suitland, Maryland, United States of America

Published on: Jan 31, 2020

DOI: <https://doi.org/10.1162/99608f92.11f3e977>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

All those interested in the success of the 2020 Census and the integrity of the federal statistical system should be grateful for Teresa Sullivan’s insightful and timely article. In this discussion, I hope to elaborate on a few themes made in the article and raise a few relatively minor quibbles.

1. Respondent Mistrust

I would like to expand on [Sullivan’s \(2020\)](#) concerns about the current climate of mistrust, outlining some of the dimensions and the links between them.

The first dimension is the respondents’ concern that their information will be intentionally disclosed by the Census Bureau. The fact that those raising this concern must go back over seven decades to find an example should serve as reassurance, and not as evidence of an imminent threat.

A related concern raised in the article is whether the Census Bureau might accidentally disclose the data, either from hacking or from a failure of disclosure avoidance. These concerns are more real inasmuch as no human system is perfect. The article mentions the loss of data by the Office of Personal Management (OPM). More damaging to the Census Bureau’s reputation was the fallout from the 2006 theft of a Department of Veterans Affairs (VA) laptop, which contained sensitive information on millions of veterans and their families. This loss by the VA led the press to ask other agencies about any laptops that they may have lost. Since the Census Bureau was an early adaptor of Computer Assisted Personal Interviewing (CAPI), it had over the years deployed hundreds of laptops without the means to update their security systems as technology changed. Further, the Census Bureau lacked an up-to-date inventory system. The charge, ‘How can we trust the Census Bureau to count people when it cannot even count its own laptops?’ was especially stinging.

In applying statistical data disclosure techniques, there is always a tradeoff between risk and the usefulness and accuracy of the published data. For example, in the 1990 Census, no disclosure avoidance techniques were applied to the race data for the redistricting (Public Law 94-171) files because of user demands. Now, the Census Bureau is developing a state-of-the-art disclosure avoidance system for all files for the 2020 Census. One might point out that the disclosure-avoidance system used for the 1960 Census was also state-of-the-art for 1960, but the data released then is at risk today. The promise is to keep the data confidential for 72 years. Can any system anticipate the technology 70 years from now so that data, especially micro-data, released now cannot be back-engineered or merged with other data sets decades in the future?

Besides external ‘hackers,’ there is a possibility of an insider threat, a Census Bureau version of Edward Snowden, the former Intelligence Community officer and whistleblower. In addition to Census Bureau employees, a large number of other people now have access to the confidential data: employees of other agencies, contractors, academic researchers, and others. These “Special Sworn Status” individuals are subject to the same criminal penalties as are Census Bureau employees. In every case, the data are shared only on a

‘need to know’ basis. Still, it is no longer possible to simply and honestly say, as we once did, ‘Only Census Bureau employees will have access to your data.’

These are risks under the control of the Census Bureau. Trust must apply to all parts of the government. The confidentiality protections of Title 13 are very strong. Still, the public has in recent years seen the Office of Legal Council (OLC) and the Foreign Intelligence Surveillance Act (FISA) courts issue secret legal guidance authorizing as permitted what would seem to have been forbidden. I strongly doubt that there is a secret legal opinion saying that the post-9/11 Patriot Act, which expanded the government’s authority for wiretapping and surveillance, allows the NSA to hack Census Bureau data streams. However, I could not offer anyone any reassurance.

Other than risk to the individual respondent, there is the question of group harm. This is codified in the American Statistical Association’s (ASA) Guidelines for the Ethical Practice of Statistics (2018):

The Ethical Statistician recognizes any statistical descriptions of groups may carry risks of stereotypes and stigmatization. Statisticians should contemplate, and be sensitive to, the manner in which information is framed to avoid disproportionate harm to vulnerable groups.

Regardless of whether the names of individual Japanese Americans were disclosed in 1942, the principal harm arose from the tabular data. Data tabulations were at a detailed geographic level, and failed to distinguish Japanese citizens from Americans of Japanese ancestry. In the mid-2000s, the Census Bureau helped the Department of Homeland Security (DHS) tabulate some data on Arabs living in this country. The data set used was publicly available, having already undergone disclosure avoidance, so the Census Bureau did nothing wrong. However, when the DHS failed to give an adequate explanation for the purpose of the tabulations, mistrust was shared with the Census Bureau.

Some of the concern about the citizenship question came from a belief that these data would help ICE locate unauthorized residents, not by identifying them individually, but by identifying group locations. That particular concern is now moot. However, the 2020 Census will be asking very detailed questions about ethnicity. The census has long asked detailed questions about Hispanic and Asian origin and ethnicity. For the first time, the 2020 Census will also be asking about European, Middle Eastern, African, and Caribbean ethnicities as part of the race question. For example, people will now be asked to report whether they are of English, Egyptian, Somali, Haitian, or Nigerian origin. Will those who feared a census that asked about their citizenship status be confident in a census asking detailed questions about their ethnicity and origin?

There is another aspect to trust: not ‘Will my response be used to harm me?’ but ‘Will my response help me?’ Decades ago, I was at a meeting with community leaders for a poor section of a large city. One said to the Census Bureau representatives, (I paraphrase) ‘You came here ten years ago and said that if we answered the census we would have better schools, better markets, better roads...our community would benefit. Well, look

around you. It's ten years later and I don't see it!' If people don't trust the government to work to improve their lives, why should they bother with the census?

What might be the implications of the lack of trust? Interestingly, the one effect that has gotten the most attention now seems to have little statistical (as opposed to political) support. That is the concern that including a citizenship question would reduce initial self-response. Until recently, all we have had were analyses based on heroic assumptions and small experiments using convenience samples. The Census Bureau has recently conducted a large-scale split-panel test. This type of study has long been the accepted gold standard to evaluate census questions. The study found neither a statistical nor an operationally meaningful difference in the overall response rates between the two panels, although some small but statistically significant differences were found for some sub-groups (Velkoff, 2019, U.S. Census Bureau 2020).

Of course, initial response is just the start of the process. There will be Non-Response Follow-Up (NRFU), including proxy response. There will be Vacant-Delete Check. There will be missing-data imputation. Importantly, the missing-data imputation includes whole-person imputation (number of people known) and count imputation (number of residents not known). How might mistrust play out throughout this process? The census begins with a high-quality list of dwelling places. Through response, proxies, or imputation, the Census Bureau will assign a count (including sometimes zero) to each of these addresses. Of course, there have always been missed and hidden units, but they have more to do with local zoning laws, and are unlikely to be a bigger problem in 2020.

There have also always been missed people in enumerated units. I believe that most of them are due to a misalignment between the respondents' conceptions of their 'real' family or household, and the complex census residence rules. However, it is certainly possible that in 2020 mistrust of government may play a bigger role in within-household misses.

Mistrust may also play out in partial responses. People who believe that the census will be used to target people of their race or ethnicity may well skip those questions. I have heard rumors that some advocates are telling people not to put down their names. In recent censuses, the minimum information to have an individual record counted has been remarkably small; just two items are required (Hogan, 2003). So 'White, Male' with no name or age was in the past considered sufficient, as was 'J.R., age 50.' It is unclear what these requirements will be for Census 2020.

Incomplete responses can have profound effects. They will make unduplication of multiple enumerations much more difficult. This includes both multiple responses for the same address, and the same person reported at multiple addresses. Incomplete responses will also make it more difficult to assess census accuracy through matching studies. Finally, respondent reluctance may lead to more whole-person and count imputations, for example, cases where all that is known is that there are three people living there or only that the unit is occupied. I discuss this issue below.

2. User Mistrust

The census must not just be accurate, it must be seen as legitimate. Roughly speaking, there are two ways to conceive a census. One way is in strictly statistical terms. Is the mean-squared error for the statistic of interest sufficiently small? It would not matter if half the people were missed and half counted twice so long as the number is correct. A 20% undercount spread evenly across all states would result in the same Congressional apportionment as a completely accurate count.

The other conception is process oriented, and almost biblical in its view of an ‘actual head-count.’ This is captured in part in the slogan “Once, Only Once and in the Right Place,” (see, for example, Jarmin, 2018). The conflict between these approaches is seen in the treatment of college students correctly counted at college, but also incorrectly counted at their parents’ home. Unduplicating these returns would no doubt increase accuracy and probably decrease the differential undercount between well-off and poorer neighborhoods. However, this unduplication, like all census processes, would not totally be without error. There could be no guarantee that no person would ever have his/her enumeration wrongly removed. So, the Census Bureau chooses to unduplicate none, leaving all in. I once had a congressional staffer almost come across the table in anger over the issue of ‘removing real people from the census.’ The more the census deviates from an actual nose count, the more it loses legitimacy from those who see the *process* as paramount.

Beyond this philosophic difference, there can be a real lack of trust of the statistician. The possibility that the administration would manipulate the results was a real concern of those who opposed statistical adjustment for the net undercount.

The risk for the 2020 Census is that the lack of trust among respondents can lead to situations that worsen the lack of trust among users. First, consider the possibility that for many households only the total number of residents is reported, or that large numbers of people refuse to give their names. This would make unduplication and fraud detection more difficult. In some ways worse, it would make it much more difficult to demonstrate that the census had not been tainted by large-scale fraud.

Secondly, consider the situation where a refusal by large numbers of households to cooperate at all leads to an increased number of count imputation, that is, the estimation of the number of people resident in a unit. These count imputations were the subject of the Supreme Court case of *Utah v. Evans* (see Cantwell, Hogan, & Styles, 2004). The issues were whether these cases constituted ‘sampling’ for the apportionment counts, and therefore unlawful, or were not ‘actual enumerations’ and therefore unconstitutional. This imputation process was judged both legal and constitutional by a thin five-member majority of the court. However, much of the ruling was based on the *de minimus* level of such imputations in the 2000 Census. With a different court and, conceivably, a higher level of count imputations, the result may be different.

What can the statistical profession do to restore trust? In the near term, the options are limited. Much of the mistrust is directed toward government in general. With the 2020 Census taking place between an

impeachment process and a presidential election, the voices of the statistical community are sure to be drowned out.

However, the profession needs to be aware of what it means to play in a highly charged political environment. The ASA guidelines make clear:

The Ethical Statistician recognizes that differences of opinion and honest error do not constitute misconduct; they warrant discussion, but not accusation.

Statisticians must always keep in mind that criticism that would be appropriate in an academic setting can be picked up and twisted in a charged political environment to impugn the integrity of the government statisticians. The Census Bureau lost much of its independence when statisticians took sides in the ‘Undercount Wars,’ my term for the bitter debates and litigation surrounding the 1980, 1990, and 2000 censuses. The default approach must always be ‘I might have done it differently, but the government statisticians’ choices are honest and reasonable.’

3. A Few Comments and Quibbles

I believe that the origins of the citizenship question lie not in a desire to destroy the 2020 Census, but are better explained by a desire to have accurate small area data on citizenship. In *Evenwel v. Abbott* (2016), the Supreme Court left open the question of whether state and local redistricting could be based on the voting eligible population. But to do this, states would need accurate small area data on citizenship by age. An *amicus* brief filed by a group of former census directors made clear that the American Community Survey (ACS) data were inadequate for the purpose.

The only voting age citizen data that exists are estimates based on a continual sampling conducted as part of the American Community Survey (“ACS”) by the Census Bureau. But ACS was not designed with redistricting in mind. The timing of ACS estimates does not align with the timing of redistricting and ACS estimates are not reported at the small geographic levels redistricts normally use to build districts. Moreover, the geographic areas at which such estimates are available carry large error margins because of the small sample sizes. These factors make the ACS an inappropriate source of data to support a constitutional rule requiring states to create districts with equal numbers of voting age citizens. (Prewitt, Groves, Farnsworth Riche, & Barabba, 2015, pp. 4–5)

The obvious solution was to add a citizenship question to the 100% counts, not to destroy the census. Now that the question has been removed, the controversy has moved to whether the Census Bureau should use administrative records to produce small-area tabulations by citizenship.

Sullivan asserts that the requirement that direct taxes also be allocated proportional to the population was eliminated by the Fourteenth Amendment. I disagree. More importantly, the Supreme Court disagreed (see

Pollock v. Farmers' Loan and Trust, 1895). Although this provision has largely lain dormant since the Sixteenth Amendment, it has recently resurfaced in connection with a proposed wealth tax.

Finally, the myth that Thomas Jefferson ‘led’ the 1790 Census is much cherished at both the University of Virginia and the U.S. Census Bureau. However, his actual role was quite modest. Congress wrote a very detailed census law, giving responsibility to the federal marshals. The marshals were appointed by the president, not the Secretary of State. Their expenses were paid by the Department of the Treasury. Jefferson’s role seems to have been limited to little more than collecting the results from the marshals and sending them on to the president.

But these are very minor points. Sullivan has provided us with an insightful article on the challenges ahead. The profession should be grateful.

Disclosure Statement

Howard Hogan has no financial or non-financial disclosures to share for this article.

References

- American Statistical Association. (2018). *Guidelines for the ethical practice of statistics*. Retrieved from <https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx>
- Cantwell, P., Hogan, H., & Styles, K. (2004). The use of statistical methods in the U.S. Census. *The American Statistician*, 58(3), 203–212. <https://doi.org/10.1198/000313004X1189>
- Evenwel v. Abbott, 578 U.S. (2006)
- Hogan, H. (2003). The accuracy and coverage evaluation: Theory and design. *Survey Methodology*, 29(2), 129–138.
- Prewitt, K., Groves, R., Farnsworth Riche, M. & Barabba, V. (2015). *Brief of former directors of the U.S. Census Bureau as amici curiae in support of appellees*. Retrieved from <https://www.scotusblog.com/wp-content/uploads/2015/10/Evenwel-FormerCensusBureauDirectorsBrief092515.pdf>
- Jarmin, R., (2018). *Counting everyone once, only once and in the right place*. United States Census Bureau. https://www.census.gov/newsroom/blogs/director/2018/11/counting_everyoneon.html
- Pollock v. Farmers' Loan and Trust Company (No. 898). Argued: March 7, 8, 11, 12, 13, 1895, Decided: April 8, 1895. <https://supreme.justia.com/cases/federal/us/157/429/>

Sullivan, T. A. (2020). Coming to our census: How social statistics underpin our democracy (and republic). *Harvard Data Science Review*, 2(1). <https://doi.org/10.1162/99608f92.c871f9e0>

United States Census Bureau. (2020). *2019 census test report*. <https://www2.census.gov/programs-surveys/decennial/2020/program-management/census-tests/2019/2019-census-test-report.pdf>

Utah v. Evans, 536 U.S. 452. (2002).

Velkoff, V. A. (2019). *2019 census test preliminary results*. United States Census Bureau. Retrieved from https://www.census.gov/newsroom/blogs/random-samplings/2019/10/2019_census_testpre.html?utm_campaign=&utm_medium=email&utm_source=govdelivery

©2020 Howard Hogan. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.