To: Policy Coordination Office, U.S. Census Bureau, Department of Commerce
From: Dan Bouk and danah boyd

November 15, 2022

Re: Soliciting Input or Suggestions on 2030 Census Preliminary Research
Docket Number 220526-0123, 2022-17647

Thank you for the opportunity to offer input into your planning for the 2030 Census. We are writing to you as individual scholars rather than representatives of any organization. We are a historian and an ethnographer who have been examining different facets of the U.S. census for the last 5+ years. In our scholarship, we regularly argue that the census is democracy's data infrastructure. So much depends on the collection, processing, and publication of census data, including but not limited to the allocation of political representation and federal funding.

At the same time, as scholars of Science and Technology Studies, we are acutely aware that infrastructures are vulnerable without maintenance. We expect that many of the responses to this Federal Register Notice will focus on the aspects of the census that are most visible to those outside the federal government – including data collection and publication. While we recognize the importance of data collection, we want to use this opportunity to publicly highlight what happens out of the public eye. In short, we want to encourage you – along with partners in civil society organizations and state and local governments – to invest in the work needed to *complete* the count and help the public understand the data they are given. The work done to process collected data, fill in the gaps, and communicate uncertainty inherent in all datasets and publications is as essential to a successful census as effective data collection.

In our remarks, we would like to encourage the bureau's 2030 planners to invest in three critical but often under-recognized aspects of the work. We also encourage all who care about the census outside the bureau to pay attention to these facets as well. The bureau should:

1. Strengthen the tools for backend processing.
2. Leverage and advance the state of the art for addressing gaps in data.
3. Embrace data uncertainty and educate stakeholders about it.

In preparing for the 2030 census, we believe that it is important both to recognize this crucial, behind-the-scenes work publicly and to publicly encourage the bureau to invest in systems development and maintenance so that democracy's data infrastructure remains robust.

Backend Processing

In the minds of the public, the census is synonymous with the data collection phase. Poems, paintings, and advertisements celebrate the census taker, not the bureaucrat in the back room who works to resolve discrepancies and so ensures a complete and accurate count. We know that backend processing involves intensive coordination of hundreds of people, dozens of operations, and a wide array of software, scripts, and technologies. This sort of work usually goes unnoticed by the public at large and even by many census advocates and policymakers. More people may have realized the significance of backend processing because of how that work in 2020 and 2021 contributed to controversial pandemic-induced delays. Still, the work remains as invisible as it is essential.

We encourage the bureau to invest as much in automating backend processes in 2030 as it did in automating components of field operations in 2020. Investments in in-office address canvassing, administrative records, internet self-response, handhelds, real-time workload allocation, and much more allowed the bureau to remain resilient and continue the work even when the pandemic upended its hiring plans, advertising campaigns, and partnership programs. However, even with new tools to do real-time data quality review, the ability to automatically feed these insights back into the field operations cost time. Many data review processes are simplistic scripts that take hours to run; they can be better optimized. Many reports are manually created to support decision-making; many of these could be automated. In 2020, technical anomalies cost the bureau significant time; the reduction of funding for end-to-end testing meant that many technical glitches were not found until the last moment. Many internal systems used in data processing and publication are patched together with duct tape and require significant labor.

We recognize that the lack of automation in backend processing did not cause the delays in releasing 2020 data; we acknowledge that 2020 planners (and re-planners) accurately accounted for the time needed to accommodate slower backend systems. However, we also recognize that when the pandemic upended the schedule, the lack of efficiency in these backend systems contributed to delays and they continue to contribute to delays in publishing 2020 census products. Moreover, the public's lack of understanding about the important work of back-end processing led to confusion and irritation with the bureau. Investing in both the technical work of these back-end systems and public understanding of what it takes to make democracy's data can ensure future resilience to unexpected challenges.

We encourage the bureau to actively review all internal software systems – from the large and complex production systems to the small single use scripts – to identify places where automation, maintenance, and development can improve these systems so that the talented civil servants who rely on them can be more effective at their jobs.

We also encourage the bureau to invest in in-house technical talent whenever possible rather than relying exclusively on contractors. As we have repeatedly witnessed in complex sociotechnical systems, relying on vendors can introduce new vulnerabilities. Proprietary technologies created by third-party software vendors can make real-time debugging difficult. Moreover, by not investing in creating a pipeline of federal technical workers, the bureau's capacity for doing and overseeing future technical work diminishes. Mission-critical systems should be built in-house or commissioned in a manner that ensures that the bureau has complete control over the code. Cost-savings should not be the only calculation used in evaluating whether to outsource; local knowledge, dedication, mission familiarity, and other hard-to-measure qualities make in-house knowledge exceptionally valuable. For much of its history, the Census Bureau drove innovation in computing and tabulating technologies. We are confident it can do so again.

Note to the Commerce Department, White House, and Congress: Investment in back-end maintenance will require significant funding early in the decade. We actively support this. Technical development and testing require time. Furthermore, while we recognize that federal *policies like OMB Circular A-76 require fiduciary evaluations between in-house systems and external vendors, we also recognize the limits to those evaluations. As we've repeatedly seen for decades across federal agencies, contracted work tends to be delivered late and above cost. Work done for the Census Bureau cannot be late. This is not simply a management problem; it's a structural problem. In short, the incentives are misaligned. We strongly encourage elected officials to work with OMB and the Census Bureau to encourage and invest in in-house development. Local knowledge and local control are crucial to organizational resilience.*

Missing Data

We know that, every decade, the Census Bureau invests significantly in new survey design protocols, outreach campaigns, and operational mechanisms to entice the entire population to voluntarily participate in the census. However, we also know that not once since 1790 has the census counted everyone in the country once, only once, and in the right location. This North Star goal is critical, but it is impossible to achieve this through data collection alone. We are grateful that the Census Bureau has repeatedly invested in leveraging statistical knowledge to improve the count. Since 1960, the bureau's deployment of imputation techniques has helped address critical gaps to improve the quality of the data and enable the production of more detailed data. Likewise, the bureau's investment in individual-level de-duplication in 2020 was critical to ensuring that duplicates did not magnify disparities.

There is much more to be done. We are aware that the bureau continues to rely on outdated techniques for addressing missing data. The modeling techniques used to remedy gaps in data in industry, academic statistics, and the international human rights field provide greater precision than the approaches that the bureau primarily uses. We respect that, since *Utah v. Evans*, the bureau has remained cautious, preferring methods that have received the Supreme Court's tacit approval to methods what would provide greater accuracy, but more legal uncertainty. However, we believe that the time has come for the bureau to invest deeply in research and development to improve the methods used for addressing weaknesses in the data.

As part of the 2030 planning, we encourage the bureau to actively engage with and seek guidance from experts in other sectors and statistical domains who use a range of techniques to address missing data. We encourage the bureau to conduct a full review of the current state of the field and publicly discuss the range of possible techniques that *could* be implemented beyond hot-deck imputation. As part of this work, we encourage the bureau to evaluate the strengths and weaknesses of these different techniques for addressing different data collection challenges, with an eye to how different techniques could help ensure data equity.

We know that there is no one "right" way to address missing data. We also know that within the professional statistics community, there are disagreements about the strengths and weaknesses of different techniques. That said, we also know that it is important for the bureau to continue building upon and contributing to the state of the art in the field of statistics. And this is an area where investment is mission-critical, especially considering growing non-response concerns in both the decennial census and other survey products. We believe that the methods chosen should stem from a new wave of research and evaluation rather than accepting the status quo. As the population becomes more diverse, this investment is critical to ensuring data equity.

We recognize that investment in new statistical techniques often creates controversies for the Census Bureau. For this reason, we believe that it is important for the public to see the range of possible statistical methods and their strengths and weaknesses. To begin to address this, we encourage the bureau to work with third parties – including perhaps the National Science Foundation or the Committee on National Statistics – to host learning fora for those outside the statistical community to learn how the statistical community approaches missing data.

Note to census stakeholders: No one invested in producing data products wants data to be missing. Given its Constitutional mandate and the importance of self-representation, it remains crucial for the Census Bureau to continue to invest in improving data collection. However, it is equally important to recognize that no matter how much the bureau invests in data collection, some data (and people) will end up being missed in the course of a count. Investing in addressing missing data better does not preclude investing in better data collection. But better options for dealing with missing data can lead to a more complete and more accurate count. To this end, we strongly encourage all who care about differential undercounts to invest in learning about the range of statistical approaches that the bureau can use to improve data quality when data are missing.

Uncertainty is a Reality

While we appreciate the confidence with which the Census Bureau announced the 2020 population as 331,449,281, we both groaned when this number was announced. This figure creates an illusion of precision that never was, never is, and never will be. We recognize that ever since the first census in 1790, it has been a tradition to make such pronouncements and publish data precise to the ones digit. Yet even 230-some years ago, officials knew that the number wasn't perfect. They just did not have tools to communicate the uncertainty attached to such a number. Now we do.

Too many people outside the bureau take census data as irrefutable facts. Given that the Latin origin of the word "data" is "the givens," this may seem to make sense, but the ramifications are profound. Given how the bureau is configured within the federal government, the bureau tends to reinforce an illusion of precision, while leaving the uncertainty to the footnotes or to later reports. Increasingly, we are watching uncertainty in scientific and statistical work be weaponized across domains. We encourage the bureau to get ahead of this.

Every decade, the Census Bureau devises many different techniques to evaluate the shortcomings of its work, both for the sake of scientific transparency and to make investments in improving its work down the line. For example, the Post-Enumeration Survey (PES) is a fantastic tool for the bureau to ask questions about the weaknesses of data coverage. We appreciate that the 2020 PES was created for scientific self-critique, even with adjustment off the table. However, many onlookers do not appreciate what the PES is and is not. Journalists and the bureau's critics repeatedly re-narrate this product as the "right" data and use it to question the validity of the census itself.

We respect the bureau's repeated reminders that these usages fail to reckon with both methodological limitations in the PES design that parallel the census and PES sampling limitations that the bureau shares but are regularly ignored. We also recognize that the bureau provides confidence intervals alongside the PES. What concerns is that the bureau is not grappling with how and why these scientific methods are poorly understood outside the bureau. We are concerned that the disconnect between the norms of sense-making within the statistical community and those of data users, advocates, and critics is being leveraged to delegitimize the work.

This dynamic is not unique to the bureau. For example, politicians regularly contort probabilistic information about weather provided by the National Oceanographic and Atmospheric Administration into binary facts about whether the hurricane is hitting a state or not. We are also living in a contemporary context in which distrust in science, medicine, and statistics are prompting people to distrust experts and make decisions that cost lives. We believe that the public's inability to make sense of measures of uncertainty is eroding trust in federal statistics and science. We also believe that if the bureau does not proactively grapple with this dynamic, it will face a yawning gap between its internal understandings of data quality and external interpretations of its data products.

What most data users want to know is: how reliable is any given statistic that affects my community or analysis? Unfortunately, this comes down to faith more than statistics. When people trust the bureau or believe that the data are "facts," they trust the data. But this is also a vulnerability. When trust in the bureau waivers – or politicians question "the facts" – the reputation of the bureau's data is tarnished, regardless of any empirical evaluation concerning quality.

As part of the planning efforts for the 2030 census, we encourage the bureau to directly address not only how it knows the strengths and limitations of its data, but how it brings the public along. Science and good government demand transparency, but what constitutes transparency is contested. The bureau believes itself to be transparent when it releases scientific reports, but stakeholders see those same messages as an act of obfuscation. This disconnect stems from divergent ways of sense-making about the world and a growing crisis in expertise.

We recognize that the bureau is operating within the norms of science to evaluate its own work. However, the bureau must recognize that the audience of these evaluations goes well beyond the scientific community. And for better or worse, many census stakeholders have not yet been socialized into an understanding of the imperfections of data. It is crucial for the bureau to continue to advance methods to evaluate the quality of its work. But the bureau must also work to bridge the gap between what it knows about its data and how the public interprets its self-evaluations.

To this end, we encourage the bureau to research different approaches for communicating statistical quality and uncertainty to the public. We encourage the bureau to convene experts in public health and science communications who have been grappling with this challenge in other domains. Rather than simply publishing the metrics used by internal subject matter experts, communication research should contribute to shaping the development of new products that can inform broader audiences without requiring translation by external subject matter experts.

We also believe that it is important that the bureau researches and develops new technical frameworks for evaluating and communicating uncertainty that accounts for the range of limitations in the data. This requires working across the statistical community to help advance how total error and uncertainty are conceptualized, measured, and evaluated. Such advances would be beneficial to the statistical community as a whole.

Note to census stakeholders, including statisticians, politicians, and other government agencies: While we encourage the bureau to research better approaches to evaluating and communicating uncertainty, we are acutely aware that the efficacy of their efforts depends on how those outside of the bureau receive them. It is critical that those who want to see the bureau be successful also invest in bridging the communication gap. This starts with embracing uncertainty as a reality to understand rather than a weapon.

Conclusion

We continue to be in awe of the Census Bureau's resilience. Conducting a census during a pandemic while faced with countless other headwinds is no easy feat. We are deeply grateful to the thousands of civil servants who came to work each day under adverse conditions, determined to produce democracy's data because you believed that failure was not an option. We cannot thank you enough. We also recognize that the creation of 2020 data was possible because of the investments that the bureau made to increase its resilience in planning for the 2020 census. We ask that you do the same for 2030. We know that improvements to data collection are possible, but we also believe that many of the greatest vulnerabilities that the bureau faces going forward are in the layers necessary to complete the census and help the public understand the work.

We thank you for this opportunity to offer comments and thank you for your service to our nation and to democracy.

Sincerely,

Dan Bouk (Colgate University), dbouk@colgate.edu

Dan Bouk

danah boyd (Microsoft Research / Georgetown University), db1537@georgetown.edu