Gender Identification and Survey Weighting: A Shifting Landscape

Brian R. Urlacher, *University of North Dakota, USA*

ABSTRACT In October 2021, the US Census Bureau piloted a new set of questions to operationalize sex and gender identity. This move follows a larger trend across the social sciences to rethink how surveys ask about sex and gender. Although this step is normatively positive, it complicates well-established protocols for weighting survey data. This article explores the likely pitfalls for survey researchers that accompany a shift in how the US Census Bureau measures gender. A preliminary empirical investigation of survey weighting indicates that using more inclusive gender categories will not negatively affect weighting metrics. Whereas the creation of a new set of US or even global best practices in measuring gender may be helpful to survey researchers, at this stage, there remain important empirical and ethical questions that are not well understood.

In August 2021, the US Census Bureau announced that it would pilot a new, two-question sequence to measure gender identity (File and Lee 2021). This pilot may reflect recent criticism of the US Census Bureau's use of a two-category sex question,[1] or it simply may lag the rethinking of how gender, sex, and identity are operationalized in survey research.

Improving our measures of sex and gender identity is important. Tate, Ledbetter, and Youssef (2013) argued that measuring sex and gender identity more accurately is beneficial to social scientific inquiry, important for countering discrimination, and necessary to improve the delivery of professional services. This step by the US Census Bureau is normatively positive. Yet, a shift in how the Bureau measures sex and gender will affect surveys using Census data for post-stratification weighting. This includes the General Social Survey (Davern et al. 2021, 7–8), the American National Election Survey (American National Election Studies 2021, 11), and The Pew Research Center's American Trends Panel (Keeter 2019, 8), as well as many others. Continuing to use US Census data for weighting will require researchers either to shift their measures of gender or find strategies that interface their existing measures of gender with new Census categories.

In summary, survey researchers may soon face a new set of methodological and ethical questions about how to weight survey data.[2] This article briefly reviews developments in how survey researchers operationalize gender. It discusses the problem of mismatched categories from the perspective of survey weighting. Finally, I provide an illustration and evaluation of survey weighting using the new Census categories.

OPERATIONALIZING GENDER

McDermott and Hatemi (2011, 90) noted that "The concept of gender, particularly as a demographic construct, actually embodies three separate but overlapping, correlated, and distinct components." They argued that the concept of gender often invokes—whether intentionally or unintentionally—biological sex, a constellation of social and cultural patterns, and sexual preference. The US Census Bureau (2020) historically measured sex (rather than gender) using a binary male/female classification. However, the Bureau instructs respondents to select the sex with which they identify. This qualification is necessary not only because of how sex and gender are intertwined but also because neither concept maps clearly into a binary classification. Yet, this US Census Bureau practice is by no means unusual. Westbrook and Saperstein (2015, 535–36) reviewed four long-running social surveys and succinctly summarized the problem as follows:

> We find that large social surveys generally conflate sex and gender and treat the resulting conceptual muddle as a starkly dichotomous, biologically fixed, and empirically obvious characteristic. This treatment is not restricted to questions recording the respondent's sex or gender; instead, it permeates the survey documents.

During the past decade, scholars from multiple disciplines argued that a more inclusive way of measuring sex, gender, and sexual orientation in survey instruments is needed. Numerous proposals have been offered for how to operationalize gender more effectively in survey design. These approaches range from offering respondents more open-ended response options (Ansara and

Dr. Brian R. Urlacher *is professor and chair of the political science and public administration department at the University of North Dakota. He can be reached at brian.urlacher@und.edu.*

© The Author(s), 2022. Published by Cambridge University Press on behalf of the American Political Science Association.

doi:10.1017/S1049096522001081

https://doi.org/10.1017/S1049096522001081 Published online by Cambridge University Press

Hegarty 2014) to using masculinity and femininity scales that allow people to place themselves on a continuum (Magliozzi, Saperstein, and Westbrook 2016). Fraser (2018) provided a comparison and contrast of four different methods for measuring gender through survey questions. This includes an open-ended option, a predefined list of categories, asking whether a respondent is transgender, and asking about a respondent's gender as well as sex assigned at birth. This fourth approach often is described as the "two-step" approach.

In summary, survey researchers may soon face a new set of methodological and ethical questions about how to weight survey data.

Tate, Ledbetter, and Youssef (2013) conducted semi-structured interviews with both transgender and cisgender individuals and reported that both groups found the two-step method intuitive. In a subsequent field test, they demonstrated that it was unlikely that there would be significant missing data resulting from a two-step approach to measuring gender identity. Indeed, their finding aligns with more recent research by Medeiros, Forest, and Öhberg (2020), who found that there was no evidence of a negative reaction to the wording of gender-inclusive questions in three experimental surveys conducted in the United States, Canada, and Sweden. Conversely, it is less well understood how willing individuals are to indicate that they are transgender and how this hesitancy might change in response to the larger social and political environment.

The US Census Bureau uses the two-step approach in the pilot study. This survey first asks, "What sex were you assigned at birth on your original birth certificate?" with the two possible options of "male" and "female." The second question, "Do you currently describe yourself as male, female, or transgender?," is a fixed-category question that also allows for the response options of "none of these" and "did not report" (File and Lee 2021).

The Household Plus Survey 3.2 data provide a first look at how Americans responded to this new measure. These results were reported weekly during a six-week period. In the aggregate, the survey reported 48.38% of the US population as having been assigned the sex of male at birth. The female sex assigned at birth percentage was 51.62%. In the "second step," an average of 0.89% identified as transgender and almost double that percentage identified as "none of these" (1.67%). An additional 1.75% of respondents opted not to answer the second question. The remaining 95.7% was distributed with 46.2% and 49.5% selecting the male and female category, respectively.

Survey researchers who adopt the US Census Bureau measurement approach may opt to do so not only because the Census categories are more inclusive but also to have consistent measurement error.

SURVEY WEIGHTING CONCERNS

From the perspective of survey weighting, the shift in Census categories creates a potential morass. The binary categorization of sex/gender had the advantage of being easy to incorporate into weighting procedures because it had a small number of categories with a sizeable percentage of cases in each category. Researchers often are advised to "collapse" a category if it contains a small percentage (i.e., typically 5% or less) of the overall sample (Battaglia, Frankel, and Link 2008; Battaglia, Hoaglin, and Frankel 2009, 4).

The logic behind this advice is that small cell counts result in greater sampling variation for subcategories. This volatility means that sample values are more likely to be substantially different from the population and thus, on average, will require larger weights to correct. This in turn inflates standard errors. Whereas efficiency is not a valid reason to stay with the current dichotomous sex/gender coding, survey researchers must be aware of the tradeoff between categorical precision and the inflation of standard errors due to weighting more categories with comparatively few observations. As a practical matter, there will be a temptation to combine the three smallest categories (i.e., "transgender," "none of these," and "did not report") into an amalgamated "other" category. Collapsing similar or proximate categories is common practice in survey weighting, but it is unclear if these categories actually are proximate or similar.

Kennedy et al. (2022, 6) considered the meaning of nonresponse in the context of a three-category sex/gender question (i.e., male, female, and neither male nor female) and concluded that choosing not to respond is similar to choosing the "neither male nor female" option. Yet, in the context of the two-step method, this may not be the correct way to think about nonresponse. Bauer et al. (2017, 2) found that the term "transgender" was unfamiliar to approximately 30% of Americans. They cautioned that confusion around terms may produce misclassification. Given the available Census categories, the "no answer" option may be a default response for a segment of the population that is uncertain about the terminology.

Survey researchers who adopt the US Census Bureau measurement approach may opt to do so not only because the Census categories are more inclusive but also to have consistent measurement error. If researchers choose not to mirror the US Census Bureau approach, they will need to carefully consider not only differences in measurement error related to nonresponse but also how to navigate mismatched categories and the wording of questions.

THE PROBLEM OF MISMATCHED WORDING OF QUESTIONS

Current survey weighting practice typically involves weighting a two-category sex/gender survey question to the US Census Bureau binary sex/gender question. Under the two-step method, however,

https://doi.org/10.1017/S1049096522001081 Published online by Cambridge University Press

the meaning of the binary sex variable is different. The "single-question" approach encouraged individuals to respond using the category with which they identified. Yet, in the two-step approach, there is no comparable question. Rather, there is a more prescriptive question asking about the respondents' sex that was assigned on their original birth certificate. These two questions almost certainly will elicit different responses from transgender and nonbinary respondents. Thus, researchers who look historically at Census data should be aware that these two binary measures of sex are not perfectly interchangeable.

It is of concern that in the Household Plus data, the Census appears to treat the two questions as measuring the same thing. The six weekly surveys are translated into population estimates and aligned with known parameters about the population. In other words, the US Census Bureau is weighting its own survey results in generating population estimates. One of the parameters that the Bureau weights its surveys against is sex—as measured with the single question asked in other Census studies. Thus, in the pilot study, the US Census Bureau used this single-question distribution to weight the sex-assigned-at-birth responses. That is, it treated sex assigned at birth as interchangeable with a respondent's reported sex/gender identity.

If the transgender and nonbinary population is equally likely to answer male as female assigned at birth, then ignoring differences in the wording of questions may not be a concern. However, previous investigations (Crissman et al. 2017; Raymond, Wilson, and McFarland 2017) suggested that the transgender population includes a larger percentage of male-to-female than female-to-male individuals. At the very least, the empirical implications of treating the two questions as interchangeable must be examined.

THE PROBLEM OF MISMATCHED CATEGORIES

One reason why it is difficult for survey researchers to shift to more gender-inclusive categories is that there currently are no analogous categories in the US Census data that could be used for weighting. This leaves researchers in an undesirable position of excluding categories from analysis or attempting to collapse nominal categories into a binary classification. For researchers who made the conscious choice to design and use a survey instrument with gender-inclusive language, neither approach is likely to be acceptable.

Alternatively, researchers may seek to reshuffle individuals across categories to facilitate weighting. Kennedy et al. (2022) discussed seven different ways of manipulating categories and data such that a survey with inclusive gender or sex categories can be aligned with the US Census Bureau's binary categorization. Their proposals range from random reassignment of individuals into male and female categories, to removing respondents who do not identify as either male or female, to recoding data as cis-male and not cis-male. Ultimately, they concluded that "there is no single good solution that can be applied to all situations." Nevertheless, if the US Census Bureau retains its binary measure of sex, social scientists likely will need to test and document the consequences of these imperfect solutions.

AN ILLUSTRATION AND EVALUATION

To illustrate the methodological challenges faced by social scientists weighting nominal survey categories that do not align, I recount the weighting approach taken for a 2020 survey of 1,300 respondents (Urlacher 2022). The survey, which focused on mental health stigma in Utah, was designed well before the inclusive US Census Bureau categories were piloted and did not use the two-step method. Rather, the survey used a three-category classification of sex/gender: male, female, and gender diverse. Consequently, mapping the Utah survey to the Household Plus categories created *both* a category mismatch and a wording of questions mismatch.

As a researcher tasked with weighting these survey data, I had to choose between excluding the individuals who selected the gender-diverse category and finding a way to work with the non-analogous categories in either the American Community Survey or the Household Plus Survey. In navigating this categorical mismatch, I sought to balance ethics, accuracy, practicality, and flexibility as recommended by Kennedy et al. (2022). In practice, this yielded two principles: (1) create as much space for gender diversity as the available categories allow; and (2) act with as much knowledge about the consequences of one's decisions as the available scholarship can provide.

Figure 1 highlights four strategies for aligning the US Census Bureau and survey gender categories. The first strategy seeks to preserve the three distinct categories captured in the Utah survey. It is important that the percentage of those responding "No Answer" was assumed overwhelmingly to consist of cis-male and cis-female respondents and was relocated proportionally to the male and female categories. Puckett et al. (2020) found that less than 3% of transgender individuals opt to not provide an answer when given gender-inclusive options. Thus, it was assumed that transgender and nonbinary individuals would be far less likely to decline the more inclusive categorization.

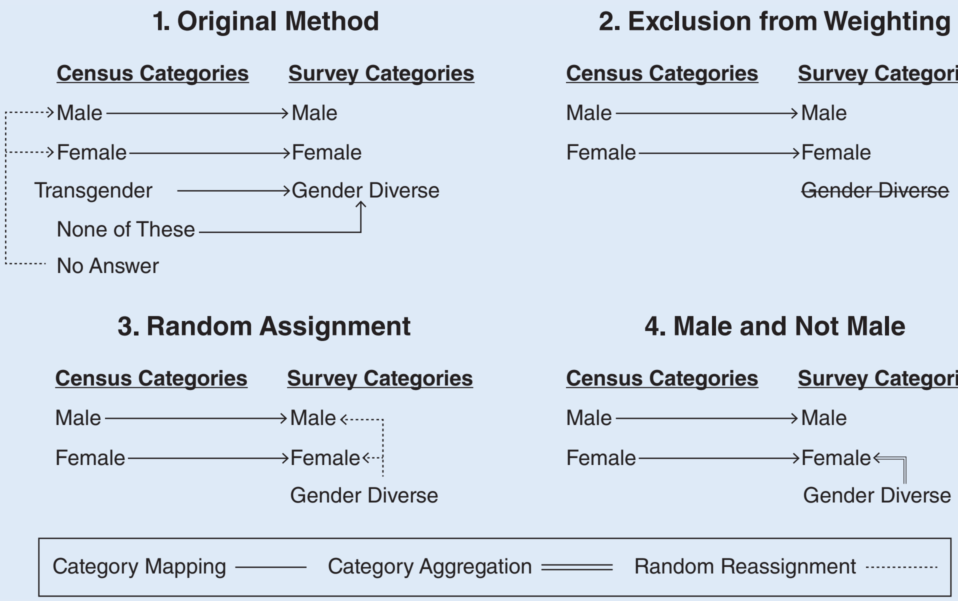
The American National Election Survey raking algorithm (Pasek 2011) was used to generate weights that would match the distribution of the original survey categories to the modified US Census Bureau categories. In addition to matching to gender categories, the data-weighting process used variables for race, ethnicity, education, age, and a geographic location (i.e., rural/urban) variable. The survey was conducted through a nonrandom process and deviated in important ways from the overall population. The original data consisted of 27.3% men (male) and 71.54% women (female). The gender-diverse category was selected by 1.15% of respondents. US Census Bureau categories for Utah were far more balanced between men and women (i.e., 48.73% and 48.59%, respectively). The collapsed Bureau categories resulted in an estimate of 2.67% for the gender-diverse population.

The raking algorithm successfully generated weights matching US Census Bureau categories and yielded a design effect (δ) of 2.64. The design effect adjusts standard errors to account for the additional uncertainty around statistical estimates that results from weighting.[3] Although it has been argued that the use of categories with a small percentage of the population risks inflating the design effect and, by extension, standard errors, the actual consequence of gender-inclusive categories on weighting has not been explored systematically.

To interrogate the effect of different categorization decisions, I reran the weighting algorithm three times, each with a different method of aggregating gender categories (see figure 1). The first re-raking of the survey simply excluded non-male or female individuals from weighting and used the original US Census Bureau sex/gender categories. This approach produced a slightly reduced design effect of 2.619. It is important to note that excluding gender-diverse respondents in the weighting stage does not

https://doi.org/10.1017/S1049096522001081 Published online by Cambridge University Press

Figure 1
Select Strategies for Category Alignment



require that they be excluded from subsequent analysis. For gender-diverse respondents, a weight of 1 still could be used. This likely would alter the alignment of percentages for other categories related to race, age, education, and geography; however, given that the number of respondents selecting the gender-diverse option in this survey is relatively small, the effect would be marginal.

An alternative approach is represented by the third option in figure 1. This approach randomly assigned "gender-diverse" individuals to either the male or the female category for purposes of weighting. Again, in the overall analysis, the three gender categories could be retained; however, for the purpose of expediency, a researcher might be tempted to lean on randomness to resolve the category mismatch. This approach yielded the smallest design effect of 2.599.

Finally, I considered an approach proposed by Kennedy et al. (2022) that takes a sociological perspective to justify a binary classification. They flagged that men have been historically socially advantaged relative to women and to transgender and nonbinary individuals. Thus, the relevant classification is the distinction between male and not male. This is the fourth option depicted in figure 1. The design effect produced in this raking procedure was 2.633.

It is important to note that across the various approaches, the change in the design effect was negligible, which suggests that the mechanical efficiency of weighting should not be a deciding factor when designing survey questions to measure gender. This analysis provides evidence counter to the argument that the status quo use of a two-category measure of gender is more efficient.

It is important to note that across the various approaches, the change in the design effect was negligible, which suggests that the mechanical efficiency of weighting should not be a deciding factor when designing survey questions to measure gender.

CONCLUSION

This article began with the observation that although there were normatively solid reasons to support a shift in how the US Census Bureau measures sex and gender identity, a shift would create challenges for survey researchers and social scientists more generally. The mismatch of categories and the wording of questions places researchers in the difficult position of justifying the collapse of categories or the recategorization of individuals. Therefore, there is tremendous value for researchers in fostering a degree of standardization. Consistent categories, over the long term, can aid in educating respondents. Confusion around terminology related to cisgender and transgender is likely to decrease as the distinction becomes more widely used.

Although this study suggests that a shift to more inclusive gender categories is unlikely to adversely affect survey weighting, there is a range of ethical and empirical considerations that remain. Whereas universal adoption of a more inclusive measure of gender would ensure more uniform measurement error across studies, we do not know what the optimal approach to measuring sex and gender actually is in terms of measurement error, theoretical salience, and empirical validity. Relatedly, we have minimal data on how stable these measures are over time. As political

https://doi.org/10.1017/S1049096522001081 Published online by Cambridge University Press

rhetoric targeting transgender people escalates, responses to questions about sex and gender provided by transgender, nonbinary, and intersex individuals may shift as well. Finally, there may be situations in which greater precision in survey questions carries risks. Even in relatively large datasets, it may be possible to analytically identify individual survey respondents.

A shift in measuring gender by the US Census Bureau will pose new challenges that researchers are just beginning to consider. In the short term, social science researchers undoubtably will muddle through these challenges. However, the more quickly that the US Census Bureau can adopt a new, empirically validated strategy for measuring sex and gender identity the better. Indeed, there are efforts underway to establish an international standard around the measurement of sex and gender through survey instruments (United Nations Economic and Social Council 2019). From an empirical perspective, such a standard would facilitate not only the convergence of the US Census Bureau and US-focused researchers on a common approach but also could facilitate cross-national research related to sex, gender, and identity.

CONFLICTS OF INTEREST

The author declares that there are no ethical issues or conflicts of interest in this research. ■

NOTES

1. The US Census Bureau classifications have been critiqued for many years (Fernandes 2017). However, the Bureau's decision to pilot new questions followed a denunciation of the binary sex/gender classification by the singer of popular music, Taylor Swift. During an event commemorating the 1969 Stonewall Uprising, Swift stated that "I got my Census the other day, and there were two choices for gender. There was male and female, and that erasure was so upsetting to me, the erasure of transgender and nonbinary people. When you don't collect information on a group of people, that means you have every excuse in the world not to support them. When you don't collect data on a community, that's a really, really brutal way of dismissing them" (King 2020).
2. This development is not limited to US Census data. Other countries (i.e., Canada, Sweden, the United Kingdom, and Nepal) have taken steps to expand census gender categories beyond a binary classification.
3. In more precise terminology, the design effect is a ratio between the variance of a weighted parameter and the variance of that same parameter if it had been drawn as a simple random sample (Kish 1965). Thus, any design effect greater than 1 will increase standard errors relative to a simple random sample of similar size.

REFERENCES

American National Election Studies. 2021. "ANES 2020 Time Series Study Full Release."https://electionstudies.org/data-center/2020-time-series-study.

Ansara, Y. Gavriel, and Peter Hegarty. 2014. "Methodologies of Misgendering: Recommendations for Reducing Cisgenderism in Psychological Research." *Feminism & Psychology* 24 (2): 259–70.

Battaglia, Michael P., Martin R. Frankel, and Michael W. Link. 2008. "Improving Standard Poststratification Techniques for Random-Digit-Dialing Telephone Surveys." *Survey Research Methods* 2 (1): 11–19.

Battaglia, Michael P., David C. Hoaglin, and Martin R. Frankel. 2009. "Practical Considerations in Raking Survey Data." *Survey Practice* 2 (5). DOI:10.29115/SP-2009-0019.

Bauer, Greta R., Jessica Braimoh, Ayden I. Scheim, and Christoffer Dharma. 2017. "Transgender-Inclusive Measures of Sex/Gender for Population Surveys: Mixed-Methods Evaluation and Recommendations." *PLoS ONE* 12 (5): 1–28.

Crissman, Halley P., Mitchell B. Berger, Louis F. Graham, and Vanessa K. Dalton. 2017. "Transgender Demographics: A Household Probability Sample of US Adults, 2014." *American Journal of Public Health* 107 (2): 213–15.

Davern, Michael, Rene Bautista, Jeremy Freese, Stephen L. Morgan, and Tom W. Smith. 2021. "*General Social Survey 2021 Cross-Section*." Chicago: National Opinion Research Center.

Fernandes, Praveen. 2017. "The Census Won't Collect LGBT Data. That's a Problem." *New York Times*, May 10. www.nytimes.com/2017/05/10/opinion/the-census-wont-collect-lgbt-data-thats-a-problem.html.

File, Thom, and Jason-Harold Lee. 2021. "*Phase 3.2 of Census Bureau Survey Questions Now Include SOGI, Child Tax Credit, COVID Vaccination of Children*." Washington, DC: US Census Bureau. www.census.gov/library/stories/2021/08/household-pulse-survey-updates-sex-question-now-asks-sexual-orientation-and-gender-identity.html.

Fraser, Gloria. 2018. "Evaluating Inclusive Gender Identity Measures for Use in Quantitative Psychological Research." *Psychology & Sexuality* 9 (4): 343–57.

Keeter, Scott. 2019. "*Growing and Improving Pew Research Center's American Trends Panel*." Washington, DC: Pew Research Center. www.pewresearch.org/methods/2019/02/27/growing-and-improving-pew-research-centers-american-trends-panel.

Kennedy, Lauren, Katharine Khanna, Daniel Simpson, and Andrew Gelman. 2022. "*He, She, They: Using Sex and Gender in Survey Adjustment*." New York: Columbia University. Unpublished manuscript, last modified March 24.

King, Ashley. 2020. "Taylor Swift Blasts US Census for Not Counting Transgender, Non-Binary People." *Digital Music News*, June 28. www.digitalmusicnews.com/2020/06/28/taylor-swift-blasts-us-census-transgender-nonbinary.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley & Sons.

Magliozzi, Devon, Aliya Saperstein, and Laurel Westbrook. 2016. "Scaling Up: Representing Gender Diversity in Survey Research." *Socius* 2:1–11.

McDermott, Rose, and Peter K. Hatemi. 2011. "Distinguishing Sex and Gender." *PS: Political Science & Politics* 44 (1): 89–92.

Medeiros, Mike, Benjamin Forest, and Patrik Öhberg. 2020. "The Case for Non-Binary Gender Questions in Surveys." *PS: Political Science & Politics* 53 (1): 128–35.

Pasek, Josh. 2011. "ANES Raking Implementation." Comprehensive R Archive Network. http://cran.r-project.org.

Puckett, Jae A., Nina C. Brown, Terra Dunn, Brian Mustanski, and Michael E. Newcomb. 2020. "Perspectives from Transgender and Gender Diverse People on How to Ask About Gender." *LGBT Health* 7 (6): 305–11.

Raymond, Henry F., Erin C. Wilson, and Willi McFarland. 2017. "Transwoman Population Size." *American Journal of Public Health* 107 (9): e12.

Tate, Charlotte C., Jay N. Ledbetter, and Cris P. Youssef. 2013. "A Two-Question Method for Assessing Gender Categories in the Social and Medical Sciences." *Journal of Sex Research* 50 (8): 767–76.

United Nations Economic and Social Council. 2019. "In-Depth Review of Measuring Gender Identity Note by Canada and the United Kingdom." *67th Plenary Session of the Conference of European Statisticians*. Paris, June 26–28. Geneva, Switzerland: Economic Commission for Europe.

Urlacher, Brian R. 2022. "Replication Data for 'Gender Identification and Survey Weighting: A Shifting Landscape.'" https://doi.org/10.7910/DVN/P6IHTM, Harvard Dataverse, Version 1.

US Census Bureau. 2020. "*Why We Ask the Sex Question*." Washington, DC: US Census Bureau. www2.census.gov/programs-surveys/decennial/2020/partners/outreach-materials/handouts/why-we-ask-the-sex-question.pdf.

Westbrook, Laurel, and Aliya Saperstein. 2015. "New Categories Are Not Enough: Rethinking the Measurement of Sex and Gender in Social Surveys." *Gender & Society* 29 (4): 534–60.

https://doi.org/10.1017/S1049096522001081 Published online by Cambridge University Press