He, she, they: Using sex and gender in survey adjustment

Lauren Kennedy*
School of Computer and Mathematical Sciences, University of Adelaide,
Katharine Khanna
Department of Sociology, University of Maryland, College Park,[†]
Daniel Simpson
.txt Engineering, New York,
Andrew Gelman
Department of Statistics and Department of Political Science,
Columbia University,
Yajun Jia
Center on Poverty and Social Policy, Columbia University
and
Julien Teitler
School of Social Work, Columbia University

May 13, 2025

Abstract

Fair and transparent handling of data is essential to reproducible and ethical statistical practices. In this work we consider the statistical challenges caused by ambiguous and imprecise measurement and use of important identity constructs. We focus on a specific example of the use of sex and gender when adjusting for survey non-response to generalization from samples to populations. This is challenging not only in that response categories differ between sex and gender measurement, but also in that both these attributes are potentially multidimensional. We reflect on similarities to measuring race/ethnicity before considering the ethical and statistical implications of the options available to us. We demonstrate how to weigh different facets of the

*The authors gratefully acknowledge the U.S. Office of Naval Research, National Science Foundation, National Institutes of Health, the Natural Sciences and Engineering Research Council of Canada, the Canadian Research Chair program and the Institute of Education Sciences for their partial support of this work.

[†]acknowledges support from the National Science Foundation Graduate Research Fellowship, grant number: DGE-1644869.

decision process with the New York City Poverty Tracker survey. We conclude not with a single recommendation for all surveys but rather with an awareness of the complexity of the problem and the benefits and limitations of different approaches.

Keywords: Data management; Measurement; Survey research

1 Introduction

The reproducibility crisis has encouraged a greater focus on transparency at all stages of the research process. By sharing analysis code, we ensure not only that analyses are reproducible but that all stages and decisions of the analysis can be scrutinized. The scrutiny of decisions encompasses those that are not just statistically impactful, but also those that reflect the respect shown to the subjects represented by the data. However, there can be competing needs—what is convenient statistically and what is most respectful to the subjects.

One example of this is the management of variables that represent subject identity. In this manuscript we deal specifically with the handling of sex and gender variables in survey data analysis but the lessons generalize to the handling of sex and gender variables more broadly in the social sciences, as well as other variables that represent measured identity constructs.

Our article proceeds as follows. First we review sex and gender constructs in the social sciences. Then we explain the particular statistical challenge faced by surveys that adjust by sex or gender. The remainder of the article we spend discussing the solutions to this challenge, with a particular emphasis on both a statistical and ethical consideration.

1.1 Measuring sex and gender

Since the 1950s, psychology has distinguished between sex and gender (see Muehlenhard & Peterson 2011, for a historical overview), with gender becoming increasingly considered more relevant for social research (Basow 2010). Glasser & Smith (2008) also cite instances where gender and sex are "vague, conflated and apparently synonymous" (p. 345), particularly when they are measured in binary terms.

We use the definitions provided by the Canadian Institutes of Health Research (2020):

> "Sex refers to a set of biological attributes in humans and animals. It is primarily

> associated with physical and physiological features including chromosomes, gene expression, hormone levels and function, and reproductive/sexual anatomy. . . .
>
> Gender refers to the socially constructed roles, behaviours, expressions and identities of girls, women, boys, men, and gender diverse people. It influences how people perceive themselves and each other, how they act and interact, and the distribution of power and resources in society. . . . "

The terms "male" and "female" are generally preferred when referring to sex, whereas "man" and "woman" are typically preferred when referring to gender, in order to distinguish between these biological and social aspects (Canadian Institutes of Health Research 2020, Muehlenhard & Peterson 2011, West & Zimmerman 1987). For most people, the question of which construct is being measured is irrelevant—their responses would directly correspond (male/man vs. female/woman). However, for the subset of people for whom sex and gender differ, we realize that both variables are multidimensional. For example, in the above definition, sex has at least four dimensions (chromosomes, gene expression, hormones, and anatomy) as does gender (roles, behaviors, expressions, and identities).

When measuring gender with simply two categories, there is a failure to capture the unique experiences of those who do not identify as either women or men, or those whose gender does not align with their sex recorded at birth. To rectify this, researchers are recommended to include a more diverse set of possible responses. Cameron & Stinson (2019) recommend open responses for gender, but for the particular problem of survey adjustment we focus on three response categories, as survey adjustment typically works with discrete variables.

Race and ethnicity, much like sex and gender, are socially constructed identities that are constituted through a range of attributes (skin color, facial features, country of birth, racial self-identification, language, and culture). Survey research on race and ethnicity has similarly grappled with the challenges of measuring these identities in meaningful and consistent ways. As Roth (2016) noted, “With the word ‘race’ used as a proxy for each of these dimensions, much of our scholarship and public discourse is actually comparing across several distinct, albeit correlated, variables” (p. 1310). The multidimensionality of identities like race and gender suggests that precisely what we measure, and precisely how this measure is interpreted by the respondent, can have profound impacts on our findings.

To account for this multidimensionality, some research has examined the advantages of employing multiple measures within a single survey to better understand the implications of each dimension for social inequality. Saperstein et al. (2016) analyzed such data and argued that “the relative importance of various dimensions of race likely depends on the outcome in question.” With these issues in mind, it remains an open question how researchers should proceed when they want to compare across data sets that inconsistently employ single measures of race (or gender). The importance of the outcome of interest in determining the best operationalization to measure inequality suggests that there may not be a one-size-fits-all solution for how to merge two such data sets.

Prior research has also shown that racial identity is not always stable over time. In a study comparing the 2000 and 2010 U.S. censuses, Liebler et al. (2017) found that change in racial self-identification is common, especially among those who do not fall squarely within the single-race White, Black, or Asian categories. Moreover, racial self-identification does not always match others’ perceptions of one’s own race. This matters not only because others’ perceptions may be important determinants of inequality, independent of racial self-identification, but also because racial contestation itself is an increasingly prevalent social process that contributes to strength of racial group commitment and identification (Vargas & Stainback 2016), which studies have shown is associated with a range of social attitudes and behaviors (Abascal 2015, Ellemers et al. 1999, 2002).

Measurement of sex and gender has similar problems, but fewer studies have examined survey items that measure sex and gender in different ways. Bittner & Goodyear-Grant (2017) note, “The principal problem is the conflation of gender with sex in survey research. Consequently, gender is typically treated as a dichotomy, with no response options for androgynous gender identities, or indeed degrees of identification with masculine or feminine identities” (p. 1019). Inconsistency in which attribute is measured (sex or gender) and the range of responses available pose challenges when combining or comparing across multiple surveys. One reason for the limited study of sex and gender minorities is that they represent a smaller fraction of the general population, compared to many racial or ethnic minorities. Properly adjusting for sex and gender classification becomes more urgent when studying targeted subgroups or when measuring low-frequency attitudes or behaviors in terms of statistical bias, but for the dignity of respondents it is always relevant.

1.2 Statistical and data management challenges

The challenges faced by surveys are increasing. Pew Research Center (2019) and others report continuing declines in response rate. This creates a potential threat for representation, as a lower response rate can increase the impact of heterogeneous response between demographics. A lower response rate also increases costs, leading to a surge in the use of non-probability samples and panel-based sampling. At the same time, surveys are often asked to do more with less data, generally in the form of creating estimates of smaller demographics and geographic areas. Political polls are on aggregate no less accurate than in decades past, but they have come under increased pressure as more has been expected of them in the busy news-media landscape (Gelman 2021).

Usual weighting or poststratification-based adjustments require demographic counts of adjustment variables to be known in the population. This is a problem when demographic questions are asked with different wording or different measurement categories compared to the census. One solution to this is to mirror the census demographic questionnaire to ensure comparable measurements. This is not always desirable or possible because rather than the survey question choice being driven by the needs of the survey, instead it would be driven by available population information.

One instance of this is the measurement and use of sex or gender in surveys. The U.S. census measures sex, but increasingly surveys have moved to measurement of gender as a key demographic variable. This creates challenges in poststratification adjustment as two different but related demographics need to be harmonized. In addition, the measurements themselves may not be analogous because while sex is typically measured as a binary category, gender can be measured with more responses or even as a continuum.

Although there may be no good way to resolve sex and gender measurement in surveys, it is an important problem both structurally and to individuals. Despite the strong overlap between the two variables, there is no simple mapping from sex to gender that works for the entire population. This is true even for binarized gender and sex but becomes especially urgent once we consider intersex, transgender, non-binary, and other categories. This affects survey research, not just for surveys that directly concern gender and sex roles, but also in non-response adjustment which is common in many surveys. For example, it has long been standard practice for political polls to adjust for sex, along with other variables such as age, ethnicity, and education (Voss et al. 1995), in part because women have traditionally responded to surveys at a higher rate than men.

For online surveys, which rely solely on poststratification to adjust an unrepresentative sample to the population of interest, the decision of how to do so will be particularly important. Kennedy & Gelman (2021) attempted to adjust a survey which was developed by psychologists. In it, gender was measured with three categories, but the U.S. census measured sex as two. For simplicity in presentation, the authors have removed those who responded “other” from the dataset. However, they note that this solution might not be appropriate. In this work we consider the problem more extensively.

It is not just psychology research that has faced this challenge. Other general population surveys have begun adding non-binary gender options. For example, in 2020, the New York City Longitudinal Survey of Wellbeing, also known as the Poverty Tracker (Collyer et al. 2020, Columbia Population Research Center 2012) recognized that their existing measurement of gender as exclusively male or female did not reflect their desire to respect respondents’ identity. They moved to offering a non-binary option in measuring gender identity. The settlement of measuring gender identity more inclusively in the Poverty Tracker came from respecting feedback from respondents’ requests of including a non-binary option as well as the desire to be inclusive and representative of non-binary identity populations. The question remains open how to appropriately adjust the sample when only sex is known at a population level. Although surveys and censuses may some day all move to using multidimensional, consistent measures of both sex and gender, adjustment will remain a perennial issue for any longitudinal data analysis, meaning that these ethical and statistical concerns are not going away anytime soon.

It is increasingly apparent that there will be no one-size-fits-all measurement solution, which means there can also be no automatic statistical solution. For example, the U.K. Office for National Statistics (2020) recommends using a second question that asks whether the respondent identifies as the same gender as their sex registered at birth, and free response if not. This is similar to the differences in race/ethnicity questions between different countries that make cross-national and cross-time research difficult.

Indeed, the population as measured in the census may not even be the target population of interest. If the population of interest has a higher proportion of non-binary people, appropriate poststratification of a sample to the population counts will potentially have a larger impact on overall and subgroup estimates.

In this paper, we consider two challenging scenarios:

1. A survey has three or more response categories to elicit gender, but we wish to poststratify to a population where sex is measured as either male or female. This will arise, for example, when raking to the U.S. census.
2. We want to combine data from multiple surveys that ask sex or gender in different ways, or allow different responses to these questions.

Unless the survey or census question is very specific, responses can capture a mix of all the dimensions of sex and gender listed above. For example, the 2020 U.S. census asks, "What is Person 1's sex? Mark ONE box: male or female." The subset of people who might have difficultly responding to this question can choose what aspect of sex or gender they would like to use in their response. Even though the variable is labeled as sex, the response can include some aspects of gender, as is there some freedom in what biological sex characteristics are used in the response.

For surveys that allow non-response, there is a difference between a respondent whose response to gender is missing, and one who actively chooses a category that is neither male or female. For missing respondents, imputation is a procedure that assigns potential values to the respondent had they responded, generally using a model or some other information and in such a way to respect uncertainty of these potential values. If the missingness is truly missing at random, then this is not particularly unethical, but it should be remembered that respondents who do not identify as male or female may choose to skip this question in protest or because they are not sure how to respond. In this case, non-response is disproportionately akin to answering as neither male or female. For those respondents who are given the option of more than two categories and have actively indicated that they do not identify as either male or female, they should not be identified as such.

Some large surveys are moving to measure both gender and sex recorded at birth. For instance, the 2021 Canadian census planned to measure sex at birth and gender identity separately (Statistics Canada 2020). The General Social Survey also began this practice in 2018 (Smith & Son 2019, Lagos & Compton 2021), although sex and gender are still confused, with responses to "What is your current gender?" referred to as `SEXNOW`. This does not resolve all challenges as sex at birth does not capture all the dimensions of biological sex, nor does the response to a gender identity question capture all dimensions of gender-related roles, behaviors, expressions, and identities. This is a recent development limited to only some countries. It is likely that this challenge will remain relevant for some time.

The move to measuring gender with (at least) three categories has highlighted an already existing problem. To adjust by gender, we must first create some sort of mapping from gender to sex. When gender was measured in binary format, naively, man could be mapped to male and woman mapped to female. The addition of a non-binary category, however, forces us to consider mapping sex to gender more generally.[1]

For survey weighting techniques where only gender is adjusted for, one temptation is to simply give the "non-binary" respondents an average weight. This avoids imputing sex or gender, but implies that the weight for non-binary respondents should be dependent on the relative ratio of the over/undersampling of male and female respondents, and there is no reason to believe this is the case. In addition, this continues to perpetuate the conflation of gender and sex. Aside from this, in many surveys as we adjust for more and more variables, we increasingly rely on methods like raking. For these methods, each category in the sample

[1] We use non-binary as a category name for those who identify as non-binary, agender, gender fluid, and other gender identities outside of woman and man. Although "other" is commonly used for this category, we specifically do not use this term to avoid othering those who do not identify as men or women.

	Impute sample	Remove respondents	Impute population
Assume population distribution	Yes	No	Yes
Model population distribution using auxiliary data	Yes	No	Yes
Estimate gender using auxiliary information	Yes	No	No
Impute all non-man as female	Yes	No	No
Remove all non-binary respondents	No	Yes	No

Table 1: *Possible options for scenario 1. Columns represent potential facets of ethical consideration, while rows represent possible facets of statistical consideration. The cells represent whether it was possible to address these considerations together.*

has to be matched to a category in the population. Similarly, for methods such as multilevel regression and poststratification (Gelman & Little 1997, Park et al. 2004), one would either need to (stochastically) impute a binary variable in the sample or else construct a model on the expanded space with three or more options.

We aim here to consider potential options available. Key concerns are *ethics* (respecting the perspectives and dignity of survey respondents), *accuracy* (for estimates of the general population and subpopulations of interest), *practicality* (using more complicated procedures only when they serve some useful function), and *flexibility* (anticipating future needs).

2 Poststratification and gender measurement

We consider a scenario that is increasingly common within the United States. A survey measures gender with three response categories (man, woman, and non-binary), but the population data to which we would like to poststratify to measures sex with two response categories (male and female). In this scenario there are multiple issues at hand. First, as we have discussed, sex and gender are separate and distinct constructs. Second, even if they were the same construct, they are measured with different potential categories. We create a matrix of potential solutions in Table 1.

One of the challenges of considering the potential options is the interaction between statistical and ethical issues. Typically, scientists are trained in either one or the other, but rarely are we educated in detail on the intersection between the two.

2.1 Ethical concerns

There are ethical concerns with the collection and protection of gender and sex in surveys. These issues include data sensitivity and security (Holzberg et al. 2017) and the purpose of collecting such information (Federal Interagency Working Group 2016). Here we assume that collected gender is necessary for adjustment and to ensure adequate representation across genders, and that security risks can be mitigated appropriately. In this section we assume that gender is measured in the sample with more than two response categories and sex is known (in aggregate) in the population.

Imputing sample sex

This method involves imputation using the gender reported by an individual in the sample to predict their potential response for their sex. In two of the three potential methods, this will involve directly imputing those who respond man as male, those who respond woman as female, and those who respond non-binary as either male or female. The remaining method (using auxiliary data) does allow the potential to impute man as female and vice versa.

As statisticians, we can impute a potential answer to a binary sex question from a non-binary gender item. However, the current confusion between sex and gender within academic literature may make it appear as if we are misidentifying their sex. In reality, we cannot know if this is the case. Although neither binary sex category directly corresponds to a non-binary gender identity, in some cases, we will correctly impute the binary sex category that respondents would have selected were they to have been asked. In other cases, we will incorrectly impute this binary category. This is of course the case with multiple imputation, as this method is inherently an imprecise estimation based on other information in the data. It is important, however, to carefully consider the ethical implications this carries, particularly because, as Keyes (2018) states, “an error rate that disproportionately falls on one population is not just an error rate it is discrimination,” or, in the words of Noble (2018), “algorithmic injustice.”

It could be argued that rather than imputing an individual’s sex, we are instead imputing their expected response to a question as posed by the census (What is your sex? M/F). This may be the methodologist’s intent, however, it is impossible to ensure that it is understood by users of the data, the survey respondents, and the populations affected by the survey analysis. In addition, this may be completed post collection without respondents’ explicit consent, which creates further ethical concerns.

Another challenge to this technique is the consideration of imputation error. Is there a difference between imputing potential sex responses based on demographic patterns compared to other less formal imputation procedures, such as identification by interviewer or complex features of other covariates collected with machine learning methodologies? Imputation by demographic proportions reinforces the statistical need to know the proportions of different cells for survey adjustment, and seems analogous to using a method such as raking to impute potential cell proportions when only margins are known. Imputation by interviewer or through complex machine learning techniques has a greater emphasis on imputing the individual, which as we have already discussed is potentially unethical and discriminatory.

Remove respondents

This method involves removing respondents who do not identify as either a man or a woman from the sample when constructing survey weights. This technique is easily communicated to respondents, data users, and the wider public. It avoids the potential misgendering issues described in the previous section on imputing sex by avoiding assigning sex altogether.

However, this method means that the responses of non-binary individuals are not counted for any analysis where the analyst wishes to make population generalizations. Participating in a study has, at a minimum, a time cost (and can potentially have other costs) that cannot be justified if non-binary respondents' data are not used. Moreover, this structural exclusion is a form of discrimination against non-binary individuals. If one purpose of surveys is to ensure equal and fair representation, then this method actively prevents non-binary respondents from having this opportunity. When it comes to population mean estimates, removing non-binary respondents is roughly equivalent to assuming that their responses would essentially be the weighted mean between male and female estimates.

Impute population values

The ethical considerations associated with imputing the population might appear to mirror that of the sample, but there are additional nuances. To understand this, consider two different scenarios.

The first scenario is a large population ($N \to \infty$) that has been summarized by a number of discrete categories such that the number of individuals who fall within each combination is labelled N_j, where N_j is also sufficiently large. We assume that each N_j can be further split into $N_{j,\mathrm{sex}=\mathrm{f}}$ and $N_{j,\mathrm{sex}=\mathrm{m}}$. In this scenario when we refer to "imputing the population," we refer to using either a model or known distributions of response to split the cell $N_{j,\mathrm{sex}=\mathrm{f}}$ into $N_{j,\mathrm{gender}=\mathrm{w}}$, $N_{j,\mathrm{gender}=\mathrm{non-binary}}$, $N_{j,\mathrm{gender}=\mathrm{m}}$ and $N_{j,\mathrm{sex}=\mathrm{m}}$ into $N_{j,\mathrm{gender}=\mathrm{w}}$, $N_{j,\mathrm{gender}=\mathrm{non-binary}}$, $N_{j,\mathrm{gender}=\mathrm{m}}$. For sufficiently large cells this does not involve imputing any particular person's gender, thus avoiding the previous misgendering challenges.

The second scenario we consider is a relatively small population where N_j contains only a small number or even one individual. Now we can no longer ignore the finite sample effects of this imputation. This raises multiple issues. The first is that it returns us to the original problem of imputing a specific person's gender rather than an abstract expectation for a cell. The second is that it becomes more difficult to split a particular cell. For instance, if a cell in the population contains only a single individual labeled as male sex, but we wish to impute their gender to poststratify a sample, it is difficult to reflect the uncertainty of their gender, due to the relative size of expectation for each potential gender option.

2.2 Statistical concerns

Although we cannot consider statistical concerns without considering also ethical concerns, we use this section to describe potential statistical options. Instead of focusing on the differences between sex and gender as constructs, we could instead focus on the difference in respondent perception when answering a male/female response question, "What is your sex?" versus a man/woman/non-binary response to the question, "What gender do you identify as?" While sex and gender are different constructs, from the perspective of matching, the difference in responses between these two questions is how they are interpreted by the respondent. By framing it this way, we can reformulate this challenge into a purely statistical challenge of measurement.

Assume known population proportions

Assuming that we need to impute population data, perhaps the simplest approach is to use auxiliary information about the estimated table of gender distribution to impute gender at the population level. To do this we would assume a certain gender distribution in the population. Without a census, we cannot know the proportions; in the example here for simplicity we set to 49% women, 49% men, and 2% non-binary. We would then use this distribution to add gender to the poststratification table. It is likely that we would split the cell $N_{j,\,\mathrm{sex}\,=\,\mathrm{f}}$ into $N_{j,\,\mathrm{gender}\,=\,\mathrm{w}}$ and $N_{j,\,\mathrm{gender}\,=\,\mathrm{non-binary}}$ and split the cell $N_{j,\,\mathrm{sex}\,=\,\mathrm{m}}$ into

$N_{j,\,\mathrm{gender}\,=\,\mathrm{non-binary}}$ and $N_{j,\mathrm{gender}\,=\,\mathrm{m}}$. There will be some bias from respondents who identify as male sex and female gender and vice versa. This procedure also does not propagate uncertainty in the imputed gender counts into the overall model or estimates.

Use auxiliary data

This method is similar to the previous method. It can be used in either the sample or the population. If used in the sample, a model predicting sex given other variables is created for each respondent and their expected sex is imputed. If used in the population, a model predicting gender given other variables is created and each poststratification cell N_j is imputed based on the expected proportion of men, women, and non-binary respondents.

Auxiliary data can be used when a separate reference data set measures both sex and gender, as well as a number of other demographics. A model is used to model either gender by sex and demographics (if imputing gender in the population) or sex by gender and demographics (if imputing sex in the sample). The benefit of this is that it simply allows for better imputation at the cell level to encompass demographic differences in gender identity but at the sample level has the same ethical challenges as simple imputation.

Other auxiliary information is available such as voice tone in a telephone interview or facial recognition software. These should not be used to infer gender unless directly related to the outcome of interest, such as perceived gender discrimination. These systems are complex and traumatic (Ahmed 2017), are frequently trans exclusive (Keyes 2018, Lagos 2019), and can have racially unbalanced error rates (Buolamwini & Gebru 2018).

Impute all non-man respondents as female

Adjusting for sex and gender is done when these constructs relate to the outcome of interest and the probability of inclusion. For outcomes that vary based on the socially ascribed meanings of sex and gender, what matters most is how these groups are treated differently in society. We therefore might expect non-binary individuals to have outcomes more similar to those of females than those of males, given that they do not benefit from a perceived traditionally masculine gender identity. Therefore, in instances where we cannot adjust for non-binary respondents separately (for reasons of sample size or data security), one reasonable option would be to combine these individuals with those who identify as female.

While intuitive from a sociological perspective, this approaches conflates the constructs of sex and gender even further. Indeed, this variable (in the sample) could be coded as "male" and "not male," with those coded as "not male" being adjusted to "female" in the census.

Remove individuals

The statistical argument is that the proportion of individuals who respond as non-binary in a survey that does not intentionally recruit from this category is very small, less than 1% according to various sources (Meerwijk & Sevelius 2017). Unless this group is very different from those who select male or female, omitting them is unlikely to make a statistical difference to population-level estimates (as we will see), but it may make more of a difference when estimating population subgroups.

Methodological decisions in practice

The Poverty Tracker was launched in 2012 by Columbia University and the Robin Hood Foundation to track poverty, hardship, and disadvantage in New York City. Since then, the study has enrolled six representative panels of adult New Yorkers and interviewed participants three to four times a year for up to six years.

In 2020 the Poverty Tracker began to include a non-binary gender response category ("Other gender"), to answer the following question: "Can you please tell me which gender you identify with: male, female, or something else?" Before that, gender response categories allowed for only Male and Female answers. Although "man" and "woman" are generally considered the preferred response categories for gender measurement to distinguish it from biological sex, retaining "male" and "female" as gender response options provided greater continuity from the older survey item. In 2024 the Poverty Tracker moved to response categories of "man" and "woman" for current and future surveys, but we deal with the 2020 survey in this example and so use the terminology used. [2]

With the new measurement categories, 55.8% respondents identified as Female ($n = 833$), 43.7% ($n = 652$) as Male, and 0.5% ($n = 7$) as Other. To ensure representation, weights were created for each wave. The process includes an adjustment made for the number of eligible adults in the household and the overlap between cellphone and landline frames. Weights were then adjusted to known population totals using raking (Lumley 2020). Typically this adjustment uses individual (age group, race and ethnicity, education level, immigrant status, housing situation, proportion of the year worked, ratio of income to the household specific poverty line) and household demographics (number of older adults, working age adults, and children). However, the new wave faces the challenge of measuring gender with more than 2 categories but needing to adjust to sex as measured in the population.

Given the relatively low numbers of non-binary respondents, it is unlikely that statistical concerns will drive the decision making. Considering the methods from an ethical perspective, the project decided that the only viable options in this case were the "Assume known population proportions" and "Remove individuals" methods. The project team decided that, given that the change in response category was specifically due to a desire to increase the inclusiveness of the study, it would not be appropriate to impute sex in the sample. In the

[2] In the 2020 sample the Poverty Tracker also included two additional frames targeting respondents with Chinese ancestry, which we omit from this analysis for simplicity.

Estimate	Condition	Female	Male	Non-binary	Overall
Poverty	Assume known population proportions	26.79% (2.46%)	20.82% (2.42%)	Redacted	23.55% (1.72%)
	Remove individuals	26.80% (2.46%)	20.85% (2.42%)	NA	23.65% (1.72%)
Severe Hardship	Assume known population proportions	37.57% (2.82%)	30.55% (2.71%)	Redacted	33.62% (2.32%)
	Remove individuals	37.62% (2.82%)	30.66% (2.71%)	NA	33.93% (1.96%)
Limiting Health	Assume known population proportions	21.65% (2.28%)	15.40% (2.03%)	Redacted	18.44% (1.53%)
	Remove individuals	21.57% (2.28%)	15.37% (2.03%)	NA	18.28% (1.52%)

Table 2: *Comparison between two methods estimating poverty, severe hardship, and work-limiting health conditions. We display to an over-precise two decimal places to show how tiny the differences are when estimating these aggregate proportions.*

rest of this section, we focus on the difference in these two methods when estimating three key outcomes for the survey: the rates of poverty, severe material hardship, and whether the respondent had a work-limiting health condition or was in self-rated poor health.

Given the low population-level information in the U.S. of gender demographics, we assumed that the population proportions were consistent with those observed in the sample.[3] Using this "impute population values" method, the estimated poverty rate was 23.55% with a standard error of 1.72%. When removing individuals who responded Other to the gender

[3] We tried to benchmark to gender population data in New York City or New York State. The limited source we found from the New York City Department of Health suggested that the rough estimation of 0.5% for the non-binary gender group in our sample was reasonable. See https://www1.nyc.gov/site/doh/about/press/pr2019/non-binary-gender-category-to-nyc-death-certificates.page and https://www.health.ny.gov/statistics/brfss/reports/docs/1806_brfss_sogi.pdf for more information.

question, the results yield a similar poverty rate of 23.65% with the same standard error. The difference between these two estimates is tiny, but in one version non-binary individuals are represented in the estimate, whereas in the other they are not.

We further looked at the subgroup poverty estimates; again, they were largely similar between the two methods; see Table 2 for details. One major difference between these methods was that when non-binary individuals are removed from the data before poststratification, estimates for the non-binary subgroup cannot be made. This is not a big factor in the decision for this panel because we chose not to release the individual estimates for the non-binary individuals due to the small cell size ($n = 7$) and data privacy concerns. However, as more panels are measured with this new response option, there is the possibility of producing an estimate for non-binary individuals in the future by collapsing panels together.

Turning our attention to other key outcomes in the survey we see similar findings. Using weights created with a gender imputed population, the results yielded a severe hardship rate estimate of 32.76% with a standard error of 2.0%. Using the weights created by removing individuals who responded in Other, the results were similar with a severe hardship estimate of 32.74% with the same standard error. Differences were also minor when estimating the proportion of New Yorkers experiencing limiting health conditions.

Comparing the subgroup estimates, we see similar findings for the differences between the two methods. We highlight that although we could not release estimates for the non binary response category, there are significant gender differences on all three outcomes with women reporting higher rates of poverty, severe hardship, and work-limiting health conditions (6%, 6%, and 7%, respectively). This suggests that gender differences are present, and perhaps dedicated efforts to increase the proportion of non-binary individuals would be merited.

The results suggest that when it comes to estimating specific outcomes for gender groups, the method of choice makes little statistical difference: differences are in the hundredth of a percentage point for population proportions. The lack of statistical differences means that we can make decisions from an ethical framework (although a different sample might warrant greater emphasis on statistical concerns). The demonstrated differences between genders on our three outcomes provide support that there is substantial gender-related heterogeneity, which motivates the end decision to use a method imputing gender in the population as it enables us to create an estimate for all three responses in future panels.

3 Looking forward

Measurement is central to science and statistics and represents a particular challenge to survey researchers and social scientists because the constructs that we measure are changing in both importance and definition over time. An appropriate measurement of a construct today might not be an appropriate measure tomorrow. Indeed, measurement in the social sciences reflects the sociological emphasis that is placed on the underlying construct. This is a challenge faced when considering the construct of race/ethnicity, but this challenge is also faced when considering sex/gender. Our challenge increases when we consider that we are not simply moving to a more diverse way of coding sex, but instead a recognition that the construct of gender, while consistent with sex recorded at birth for many, is a different construct for others. This distinction led us to frame our methods in terms of imputing one construct from the other.

This manuscript grapples with the complexities of moving from measuring sex to measuring gender in social surveys. We do not, however, make broad recommendations for one best way to measure sex or gender, nor do we offer or a singular technique to account for measuring gender in a survey when the population measures sex. Instead, we try to consider the ethical and statistical implications of a variety of different approaches.

Constructing our argument in this way is necessary, as there is no single good solution that can be applied to all situations. There are tradeoffs between ethical and statistical concerns, and the most appropriate decision will reflect this. That said, we have argued that first and foremost in this decision should be respect and consideration for the survey respondent, followed by the ease of describing the statistical method to non-technical respondents and concerns surrounding fair representation and statistical bias.

Enumerating the potential options to Scenario 1 in Table 1, statistical and ethical concerns intersect. While specific to the challenges of measuring sex and gender, our review of these approaches, with their various advantages and tradeoffs, may be useful in grappling with the measurement of other social constructs as well. Even as measurements of sex and gender improve and become more standardized, challenges will remain in longitudinal analyses. Our hope is that this is a useful resource to guide decision making for survey statisticians and survey administrators alike.

4 Disclosure statement

The authors have no conflicts of interest to declare.

5 Data Availability Statement

Deidentified data for the Poverty Tracker are available at https://povertycenter.columbia.edu/poverty-tracker-data, but due to privacy concerns these data are insufficient to replicate the results presented in this analysis.

References

Abascal, M. (2015), ‘Us and them: Black-White relations in the wake of Hispanic population growth’, *American Sociological Review* **80**(4), 789–813.

Ahmed, A. A. (2017), ‘Trans competent interaction design: A qualitative study on voice, identity, and technology’, *Interacting with Computers* **30**(1), 53–71.

Basow, S. A. (2010), ‘Changes in psychology of women and psychology of gender textbooks (1975–2010)’, *Sex Roles* **62**(3-4), 151–152.

Bittner, A. & Goodyear-Grant, E. (2017), ‘Sex isn’t gender: Reforming concepts and measurements in the study of public opinion’, *Political Behavior* **39**(4), 1019–1041.

Buolamwini, J. & Gebru, T. (2018), Gender shades: Intersectional accuracy disparities in commercial gender classification, *in* ‘ACM Conference on Fairness, Accountability and Transparency’, pp. 77–91.

Cameron, J. J. & Stinson, D. A. (2019), ‘Gender (mis) measurement: Guidelines for respecting gender diversity in psychological research’, *Social and Personality Psychology Compass* **13**(11), e12506.

Canadian Institutes of Health Research (2020), ‘What is gender? What is sex?’, https://cihr-irsc.gc.ca/e/48642.html.

Collyer, S., Matthew, M., Bushman-Copp, L., Garfinkel, I., Kennedy, L., Neckerman, K., Teitler, J., Waldfoger, J. & Wimer, C. (2020), *The State of Poverty and Disadvantages in New York City*, Center on Poverty and Social Policy, Columbia University.

Columbia Population Research Center (2012), ‘New York City Longitudinal Survey of Well-being’, https://cprc.columbia.edu/content/new-york-city-longitudinal-survey-wellbeing.

Ellemers, N., Spears, R. & Doosje, B. (2002), ‘Self and social identity’, *Annual Review of Psychology* **53**(1), 161–186.

Ellemers, N., Spears, R. & Doosje, B., eds (1999), *Social Identity: Context, Commitment, Content*, Blackwell Publishers: Oxford, U.K.

Federal Interagency Working Group (2016), ‘Current measures of sexual orientation and gender identity in Federal surveys’, https://nces.ed.gov/FCSM/pdf/buda5.pdf.

Gelman, A. (2021), ‘Failure and success in political polling and election forecasting’, *Statistics and Public Policy* **8**, 67–72.

Gelman, A. & Little, T. C. (1997), ‘Poststratification into many categories using hierarchical logistic regression’, *Survey Methodology* **23**, 127–135.

Glasser, H. M. & Smith, J. P. (2008), ‘On the vague meaning of “gender” in education research: The problem, its sources, and recommendations for practice’, *Educational Researcher* **37**(6), 343–350.

Holzberg, J., Ellis, R., Virgile, M., Nelson, D., Edgar, J., Phipps, P. & Kaplan, R. (2017), ‘Assessing the feasibility of asking about gender identity in the Current Population Survey. results from focus groups with members of the transgender population’, https://www.bls.gov/osmr/research-papers/2017/pdf/st170200.pdf.

Kennedy, L. & Gelman, A. (2021), ‘Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample’, *Psychological Methods* **26**, 547–558.

Keyes, O. (2018), ‘The misgendering machines: Trans/HCI implications of automatic gender recognition’, *Proceedings of ACM on Human-Computer Interaction* **2**(CSCW).

Lagos, D. (2019), ‘Hearing gender: Voice-based gender classification processes and transgender health inequality’, *American Sociological Review* **84**(5), 801–827.

Lagos, D. & Compton, D. (2021), ‘Evaluating the use of a two-step gender identity measure in the 2018 General Social Survey’, *Demography* **58**(2), 763–772.

Liebler, C. A., Porter, S. R., Fernandez, L. E., Noon, J. M. & Ennis, S. R. (2017), ‘America’s churning races: Race and ethnicity response changes between Census 2000 and the 2010 Census', *Demography* **54**(1), 259–284.

Lumley, T. (2020), 'survey: Analysis of complex survey samples'. R package version 4.0.

Meerwijk, E. L. & Sevelius, J. M. (2017), 'Transgender population size in the United States: A meta-regression of population-based probability samples', *American Journal of Public Health* **107**, e1–e8.

Muehlenhard, C. L. & Peterson, Z. D. (2011), 'Distinguishing between sex and gender: History, current conceptualizations, and implications', *Sex Roles* **64**(11-12), 791–803.

Noble, S. U. (2018), *Algorithms of Oppression: How Search Engines Reinforce Racism*, NYU Press.

Park, D. K., Gelman, A. & Bafumi, J. (2004), 'Bayesian multilevel estimation with post-stratification: State-level estimates from national polls', *Political Analysis* **12**, 375–385.

Pew Research Center (2019), 'Response rates in telephone surveys have resumed their decline', https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/.

Roth, W. D. (2016), 'The multiple dimensions of race', *Ethnic and Racial Studies* **39**(8), 1310–1338.

Saperstein, A., Penner, A. M. & Kizer, J. M. (2016), 'Making the most of multiple measures: Disentangling the effects of different dimensions of race in survey research', *American Behavioral Scientist* **60**, 519–537.

Smith, T. W. & Son, J. (2019), 'Transgender and alternative gender measurement on the 2018 General Social Survey (GSS Methodology Report No. 129)', *National Opinion Research Center* .

Statistics Canada (2020), Sex at birth and gender: Technical report on changes for the 2021 Census, Technical report.

U.K. Office for National Statistics (2020), 'Sex and gender identity question development for census 2021', www.ons.gov.uk/census/censustransformationprogramme/questiondevelopment/sexandgenderidentityquestiondevelopmentforcensus2021.

Vargas, N. & Stainback, K. (2016), 'Documenting contested racial identities among self-identified Latina/os, Asians, Blacks, and Whites', *American Behavioral Scientist* **60**(4), 442–464.

Voss, S., Gelman, A. & King, G. (1995), 'Preelection survey methodology: Details from eight polling organizations, 1988 and 1992', *Public Opinion Quarterly* **59**(1), 98–132.

West, C. & Zimmerman, D. H. (1987), 'Doing gender', *Gender & Society* **1**(2), 125–151.