Using sex and gender in survey adjustment

Lauren Kennedy* Katharine Khanna† Daniel Simpson‡
Andrew Gelman§

October 1, 2020

Abstract

Accounting for sex and gender characteristics is a complex, structural challenge in social science research. While other methodology papers consider issues surrounding appropriate measurement, we consider how gender and sex impact adjustments for non-response patterns in sampling and survey estimates. We consider the problem of survey adjustment arising from the recent push toward measuring sex or gender as a non-binary construct. This is challenging not only in that response categories differ between sex and gender measurement, but also in that both of these attributes are potentially multidimensional. In this manuscript we reflect on similarities to measuring race/ethnicity before considering the ethical and statistical implications of the options available to us. We do not conclude with a single best recommendation but rather an awareness of the complexity of the issues surrounding this challenge and the benefits and weaknesses of different approaches.

1 Introduction

There may be no good way to resolve sex and gender measurement in surveys. However, it is an important problem both structurally and to individuals. While the measurements might seem similar, there is no simple mapping from sex to gender that works for the entire population. This is true for binarized gender and sex categories but becomes especially urgent once we consider intersex, transgender, nonbinary, and other categories. This affects

*Department of Econometrics and Business Statistics, Monash University

†Department of Sociology, Columbia University

‡Department of Statistical Sciences, University of Toronto.

§Department of Statistics and Department of Political Science, Columbia University, New York.

We thank 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. KK would like to acknowledge support from the National Science Foundation Graduate Research Fellowship, grant number: DGE-1644869.

Corresponding author. Email: Lauren.Kennedy1@monash.com

survey research, not just for surveys that directly concern gender and sex roles, but also in nonresponse 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. Weighting typically corrects for nonresponse patterns by doing some version of raking or poststratification to correct for differences in the sex distribution of the sample compared to some known population such as from a national census.

1.1 Changes in definitions and measurements of sex and gender

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

However, when measuring gender with simply two categories, there is a failure to capture the unique experiences of those who do not identify as either male or female, or for those whose gender does not align with their sex classification. To rectify this, it is recommended that researchers include a more diverse set of possible responses. Cameron and Stinson [2019] recommend open responses for gender, but for the particular problem of survey adjustment we focus on three response categories, as survey adjustment has always concerned itself with discrete variables.

The move to measuring gender with (at least) three categories has highlighted an already existing problem. Sex as a binary male/female response is measured by the US Census [United States Census Bureau, 2020], while gender is increasingly being measured in surveys. To adjust by gender, we must first create some sort of mapping from gender to sex. When gender was measured in binary format, male could be mapped to male, while female mapped to female (again highlighting Glasser and Smith III [2008]'s suggestion of synonymous use). The addition of a non-binary category[1], however, forces us to consider mapping sex to gender more generally. Although some census bureaus are beginning to measure both gender identity and sex assigned at birth [Statistics Canada, 2020], this is a recent development limited to only some countries. It is likely that this challenge will remain relevant for some time.

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 has to be matched to a category in the population. Similarly, for methods such as multilevel regression and poststratification [Gelman and Little, 1997, Park et al., 2004], one would

[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 female and male. Although other is commonly used, we specifically do not use this term to avoid othering those who do not identify as male or female.

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.

In this manuscript, our aims are to consider potential options available. Key concerns are *ethics* (respecting the perspectives and dignity of individual survey respondents), *accuracy* (for estimates of the general population and for subpopulations of interest), *practicality* (using more complicated procedures only when they serve some useful function), and *flexibility* (anticipating future needs). In addition, the effort spent thinking through the coding of sex and gender can be useful when considering other survey responses that involve complex measurement (such as race/ethnicity and religion).

1.2 Implications

For online surveys, which rely solely on poststratification to adjust an unrepresentative sample to the population of interest, this decision will be particularly of importance. In Kennedy and Gelman [2020], we (LK and AG) attempted to adjust an online survey to the U.S. population. The survey, which was developed by psychologists, measured gender with three categories, but the U.S. census measured sex as two. We removed those who responded other from the dataset for pedagogical reasons. However, this began our awareness of this problem and of the inappropriateness of the solution we had used. We noted this in the cited manuscript and began the more considered investigation presented here.

Of course it is not simply psychological research that has faced this challenge. 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, 2012-2020] recognized that their existing measurement of gender (M/F) didn't reflect their desire to respect respondents' identity. They moved to measuring gender with three categories (male, female, and an open-ended other option), which again raised the issue of how to appropriately weight the sample when only sex is known at a population level.

Moreover, increasingly it is becoming apparent that there will be no one size fits all measurement solution, which means there can also be no one size fits all statistical solution. For example, the Canadian census bureau has moved to measuring sex assigned at birth and gender identity in two separate questions in the 2021 census, while the UK office for national statistics [for National Statistics, 2020] is recommending the use of a second question that asks whether the respondent identifies as the same gender as sex registered at birth, and free response if not. This is not dissimilar to the differences in race/ethnicity questions between different countries, but will 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 is a community of LGBTQIA+ individuals, then it is likely that there is a higher proportion of non-binary individuals when compared to the population of the U.S. In this case, appropriate poststratification of a sample to the population counts will potentially have a larger impact on overall and subgroup estimates.

In considering this, it is worth spending a moment considering why we adjust for sex and/or gender in surveys at all. These reasons represent a myriad of sociological and statistical concerns. Sociological concerns range from historic under-representation, to current response patterns, to ensuring that those who are discriminated against and have lower power in society are represented. Statistical concerns reflect on the relationship between sex/gender and many outcomes of interest.

1.3 Definitions of sex and gender

Gender and sex have been defined in different ways at different times. 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. ...

For most people, the question of which construct is being measured is moot—they would respond the same regardless. 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).

In this paper, we consider two challenging scenarios:

1. A survey has three or more response categories to elicit gender, but we wish to post-stratify 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.

These problems are not unique to sex and gender. The first scenario, for example, arises for other survey weighting variables such as race or ethnicity, while the second scenario arises for variables such as income, which can be measured and constructed in different ways in different surveys.

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.

Some large surveys are moving to measure both gender and sex assigned at birth. For instance, the Canadian census is planning to measure sex at birth and gender identity separately in the 2021 census [Statistics Canada, 2020]. The General Social Survey also began this practice in 2018 [Smith and Son, 2019], although sex and gender are still confused, with responses to “What is your current gender?” referred to as “SEXNOW” (pg 2). This does not resolve all challenges, though, even for surveys conducted in Canada, 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, behaviours, expressions, and identities, such that measurement differences can still occur. Yet by actively measuring both sex and gender constructs, the Canadian census makes adjustment considerably simpler.

1.4 Measuring identification in survey research: Studies of race/ethnicity and sex/gender

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] notes, "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" (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 implication of each dimension for social inequality. For example, by using multiple measures of race within the same survey to disentangle what each dimension represents and how identification across the measures differs, Saperstein et al. [2016] have 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] find 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 and 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 encountered similar problems, but fewer studies have examined the consequences of survey items that measure sex and gender in different ways. Bittner and 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" (1019). Inconsistency in both which attribute is measured (sex vs. gender) and the range of responses available poses challenges for researchers who want to combine or compare 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 relevant 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.

1.5 Missingness versus non-binary genders

Imputation of demographic variables is not unusual in surveys, and can be necessary to create survey weights. 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. We also need to account for transgender people who choose a male or female gender response that is the opposite of their recorded sex. 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 that two categories, it is very clear that they have actively indicated that they do not identify as either male or female, and so they shouldn't be identified as such.

2 Scenario 1

The first scenario we consider is increasingly common within the United States. A survey measures gender with three response categories (male, female, and non-binary), but the population data that we would like to poststratify to measures sex with two response categories (male and female). In this scenario there are multiple issues at hand. Firstly, as we've discussed, sex and gender are separate and distinct constructs. Secondly, 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. In this manuscript, we are interested in both, and so we discuss the potential options first in terms of their potential ethical considerations before considering the statistical considerations.

	Impute sample values	Remove respondents	Impute population
Assume population distribution	Y	N	Y
Model population distribution using auxiliary data	Y	N	Y
Estimate gender using auxiliary information	Y	N	N
Impute all non-male as female	Y	N	N
Remove all non-binary respondents	N	Y	N

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.

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 [on Improving Measurement of Sexual Orientation and in Federal Surveys, 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.

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 male as male, those who respond female 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 female sex as male gender and vice versa.

As researchers we can be clear that we are imputing 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 makes this practice appear as statistical misgendering, where an individual is incorrectly referred to as a gender with which they do not identify [Merriam Webster, 2020]. The right to self-identify is protected by law in some jurisdictions [Ontario Human Rights Commission, 2000], with misgendering identified as a form of discrimination. In this scenario respondents have explicitly identified as neither male nor female, yet in our statistical analysis we are assigning them to this male/female dichotomy. In survey research as in other aspects of life, people have the right to define how they are identified. This continues to be true when multiple imputation is used to represent the uncertainty in the classification because, as Keyes [2018] states, "an error rate that disproportionately falls on one population is not just an error rate it is discrimination". It is algorithmic injustice [Noble, 2018].

It could be argued that rather than imputing an individual's gender, 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 to imputing potential sex responses based on demographic patterns when 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 male or female 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 others costs) that cannot be justified if non-binary respondents' data is 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 three different scenarios.

The first scenario is a very large population ($N \to \infty$) that has been summarised 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 sufficient 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 using either a model or known distributions of response to split the cell $N_{j,\mathrm{sex}=\mathrm{f}}$ into $N_{j,\mathrm{gender}=\mathrm{f}}$, $N_{j,\mathrm{gender}=\mathrm{non-binary}}$, $N_{j,\mathrm{gender}=\mathrm{m}}$ and $N_{j,\mathrm{sex}=\mathrm{m}}$ into $N_{j,\mathrm{gender}=\mathrm{f}}$,

$N_{j,\text{gender}=\text{non-binary}}$, $N_{j,\text{gender}=\text{m}}$. A simpler version would be to split the cell $N_{j,\text{sex}=\text{f}}$ into $N_{j,\text{gender}=\text{f}}$, $N_{j,\text{gender}=\text{non-binary}}$ and $N_{j,\text{sex}=\text{m}}$ into $N_{j,\text{gender}=\text{non-binary}}$, $N_{j,\text{gender}=\text{m}}$. For sufficiently large cells it is clear that this does not involve imputing any particular person's gender, which avoids the previous misgendering challenges.

The second scenario we consider is a relatively small population such that N_j contains only a small number or even one individual. Unlike in the previous scenario, 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 in the population a cell contains only a single individual labelled as male sex, but we wish to impute their gender, it is difficult to reflect the uncertainty of their gender, due to the relative size of expectation for each potential gender option. A third issue is that the potential error rate is difficult to calculate with a small proportion of the population.

2.2 Statistical concerns

Although we cannot consider statistical concerns without considering also ethical concerns, we use this section to describe potential statistical options.

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, but we estimate at 49% female, 49% male, and 2% non-binary in this manuscript). We would then use this distribution to add gender to the poststratification table. It's likely that we would split the cell $N_{j,\,\text{sex}=\text{f}}$ into $N_{j,\,\text{gender}=\text{f}}$, $N_{j,\,\text{gender}=\text{non-binary}}$ and $N_{j,\,\text{sex}=\text{m}}$ into $N_{j,\,\text{gender}=\text{non-binary}}$, $N_{j,\text{gender}=\text{m}}$. There will be some bias from respondents who identify as male sex and female gender and vice versa. This procedure also doesn't allow any uncertainty in the imputed gender counts to be included in 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 and for each participant 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 male, female, and non-binary respondents.

One use case for auxiliary data is if a completely separate reference auxiliary data set measures both sex and gender, as well as a number of other demographics. This 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. This is not imputing sex or gender in the auxiliary data, but rather modelling the conditional relationship between sex given gender and other demographics or gender given sex and other demographics.

There is also other auxiliary information that 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. These systems are complex and traumatic [Ahmed, 2017], are frequently trans exclusive [Keyes, 2018, Lagos, 2019], or have racially unbalanced error rates [Buolamwini and Gebru, 2018].

Impute all non-male respondents as female

The rationale behind this method is that when we consider adjusting for gender or sex in surveys, we are really adjusting to ensure adequate representation of genders (traditionally female) that are systematically discriminated against through societal structures. Through this rationale, adjusting for those who identify as non-binary is important because it is likely that those who identify as non-binary experience at least as much oppression as those who identify as female. However, if we can't adjust for this group separately (for reasons of sample size or data security), then it seems that the next most reasonable option would to be to combine these individuals with the next most oppressed group, with whom they would likely have the most similar outcomes.

While intuitive from a sociological perspective, this approaches confuses sex and gender as constructs even further. It assumes that those who answer the census question on sex are instead answering a question on how they are perceived and treated in their society. However, while those who respond non-binary are being coded as female, it is more transparent that this isn't a case of imputing their gender as female, but rather collapsing female and non-binary into one group. Indeed this variable (in the sample) could be coded "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; see Meerwijk and 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.

Comparison of methods

To explore the differences between these seven different methods, we conduct a simulation study. Instead of focusing on the differences between sex and gender as constructs, we focus on the difference in respondent perception when answering a male-female response question, "What is your sex," versus a male-female-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.

To this end we create a simulation strategy where we simulate a population where proportions (p_m, p_f, p_o) respond male-female-non-binary when asked what gender they identify with. We then assume that a portion who respond male/female to this question also respond male/female when asked what sex they are. For those who respond non-binary to the first question, we simulate that they respond male to the second question from 0% (no one) to 100% (everyone), which is represented by the x-axis on Figure 1. A summary of the response patterns is given in Table 3. Studies investigating the difference in response to these questions have estimated that less than 1% of the population would respond as non-binary when given the option [Meerwijk and Sevelius, 2017]. However, this would depend on both the question framing (see the much higher proportions when gender is expressed as a continuum) and option availability (e.g., the Australian census required respondents to request the non-binary category rather than simply offering it). In our simulation study we use population gender proportions of 49% female, 49% male, and 2% non-binary.

For each scenario we simulate a sample of size 500. We then imagine potential response strategies to a question eliciting binary sex given responses to gender. We assume that a small proportion (2%) of those who respond female to the gender question would respond male to a sex question, and a similar proportion (2%) of those who respond male to the gender question would respond female to the sex question. We then vary the proportion of those who respond non-binary to the gender question who would respond male to a sex question from 0 to 100% (see table 3 for enumeration of the possibilities).

We then simulate an outcome variable y such that

$$y_i \sim \text{normal}(\mu_i, 1) \text{ where}$$

$$\mu_i = \begin{cases} \mu_{\text{m}} & \text{if gender[i] is male} \\ \mu_{\text{f}} & \text{if gender[i] is female} \\ \mu_{\text{nb}} & \text{if gender[i] is non-binary.} \end{cases}$$

We then simulate four different hypothetical conditions:

1. There are no differences between gender in the outcome, $\mu_{\text{m}} = \mu_{\text{f}} = \mu_{\text{nb}}$.
2. Those who identify as non-binary have different outcomes when compared to those who identify as male and female, $\mu_{\text{m}} = \mu_{\text{f}}; \mu_{\text{m}} \neq \mu_{\text{nb}}$.
3. Those who identify as female or non-binary are different in outcome from those who identify as male, $\mu_{\text{m}} \neq \mu_{\text{nb}}; \mu_{\text{nb}} = \mu_{\text{f}}$.
4. There are different outcomes for respondents who identify as different genders, $\mu_{\text{m}} \neq \mu_{\text{f}} \neq \mu_{\text{nb}}$.

We then compare our seven potential options for adjusting male-female-non-binary measurements in the sample to a male-female measurement in the population.

1. Impute all who identify as non-binary as male or female with a 50:50 split in the sample.

Condition	Male μ	Female μ	non-binary μ
All same	0	0	0
Male, female same	10	10	0
Female, non-binary same	10	0	0
All different	10	-10	0

Table 2: *Simulations of μ for each gender for each of the four conditions for the first simulation scenario.*

	Male sex	Female sex
Male gender	47%	2%
Female gender	2%	47%
Non-binary gender	$p\times2\%$	$p\times2\%$

Table 3: *Population distribution of male, female and non-binary genders relative to sex in simulation study. The proportion of respondents with a non-binary gender who select male or female sex is varied from all male to all female in the simulation.*

2. Impute all who identify as non-binary as female.
3. Impute sex by gender information using a model in the sample:
   (a) Simulate a best case model (imputes male as male/female in the correct proportions and vice versa, and non-binary relative to the proportion who choose to respond m/f).
   (b) Simulate a worst case model (imputes male as male/female in the opposite proportions and vice versa, and non-binary opposite to the proportion who choose to respond m/f).
4. Impute gender × sex information using a model in the population:
   (a) Simulate a best case model (imputes male, female and non-binary counts in the correct proportions).
   (b) Simulate a worst case model (imputes male, female and non-binary counts in the opposite proportions).
5. Remove those who respond as non-binary and map male gender to male sex and female gender to female sex.

Then, either we use the imputed sex counts to create simple poststratification weights for the sample and then use a weighted average of the sample to estimate the population mean, or we use the imputed gender counts to create simple poststratification weights for the sample and then use a weighted average of the sample to estimate the population mean.

Estimating the population mean

The first statistic that we consider is the population average. In Figure 1 we plot the mean square error calculated over 500 iterations of samples simulated from each population type.

As we would expect, if there are no gender differences in y then the strategy chosen makes little difference (top left panel). If those who respond non-binary are different from those who respond male and female, then removing them is the worst decision to make. Imputing gender in the population produces less error than imputing sex in the sample in this case (top right panel). In the condition where female and non-binary have the same expected value then again, imputing the population gender produces the lowest mean squared error (MSE) when using an accurate model, and imputing gender in the population by proportion works as well when those who respond are equally likely to identify as male or female sex. When all genders have different outcomes (bottom right panel), again we see that imputing gender in the population has the lowest MSE. We see a U-shaped MSE for removing those who respond non-binary and estimating gender by population proportion, likely due to simulating the non-binary gender as a value midway between male and female.

Estimating specific sex means

One reason why we poststratify by sex/gender is to ensure that when we estimate the experiences of different people of different sexes or genders, they are representative of these people. We consider in Figure 2 how the different imputation strategies impact our estimation of male and female sex, using the same simulation strategy as before. Like before, when all genders have the same expectation for y, all methods perform equally well (top row). When male and female respond the same but non-binary respond differently, we see that removing those who respond non-binary produces equally low MSE (with coding all non-binary as female having an equivalent effect when estimating male sex). Of the other conditions, removing those who respond non-binary produces the best weighted estimates, while imputing sex using the population proportions is the best method of those that do not remove those who respond non-binary. Of note is that population imputation methods imply that we cannot obtain estimates by sex, because we never estimate potential sex response in the sample.

When it comes to estimating specific gender means, the method of choice makes little difference to the mean square error; however, if those who choose non-binary are omitted, then no estimate for non-binary can be created.

3 Scenario 2

The other scenario we face is we begin to consider how longitudinal surveys should be used where measurement has changed from ”What is your sex? (M/F)” to ”What is your gender? (M/F/NB).” In this case the two measurements need to be harmonized so we can use a single sex/gender variable in a given model. The potential options will not be disimilar to those that were described above, but the survey administrators will need to carefully consider whether sex or gender is of interest. We believe that in many cases gender, as a social construction, is of greater interest to a survey researcher because it reflects the experience of people as they move throughout the world. This seems reflected in the decision of many surveys to move from measuring sex to measuring gender. In this case, gender should be imputed over survey waves where sex was measured. If the variable of interest is truly sex as

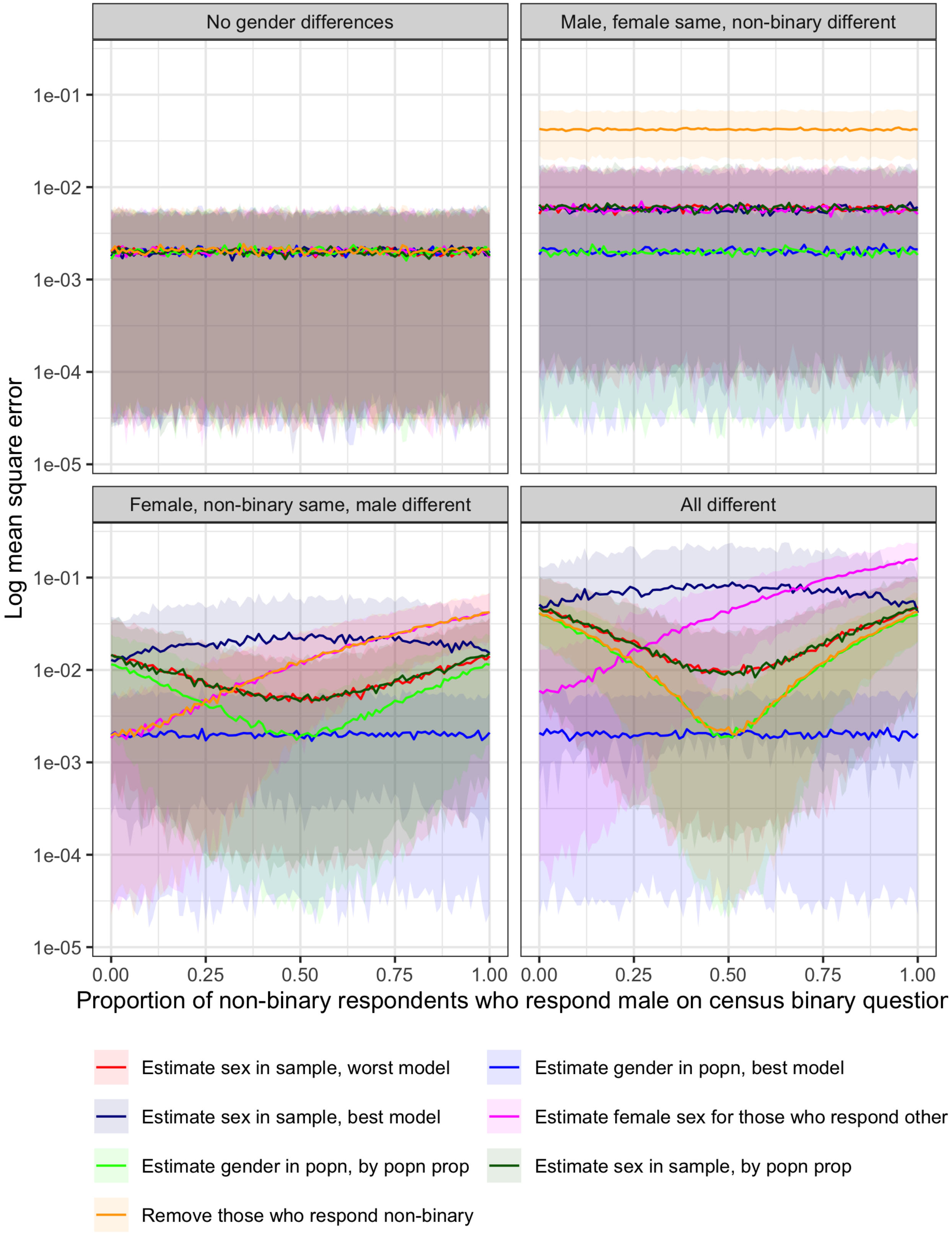

Figure 1: *Mean square difference between the true population estimate and the weighted estimate from the sample where the weights are created using the imputed gender variables. A lower mean square error is better. The colours represent different imputation strategies.*

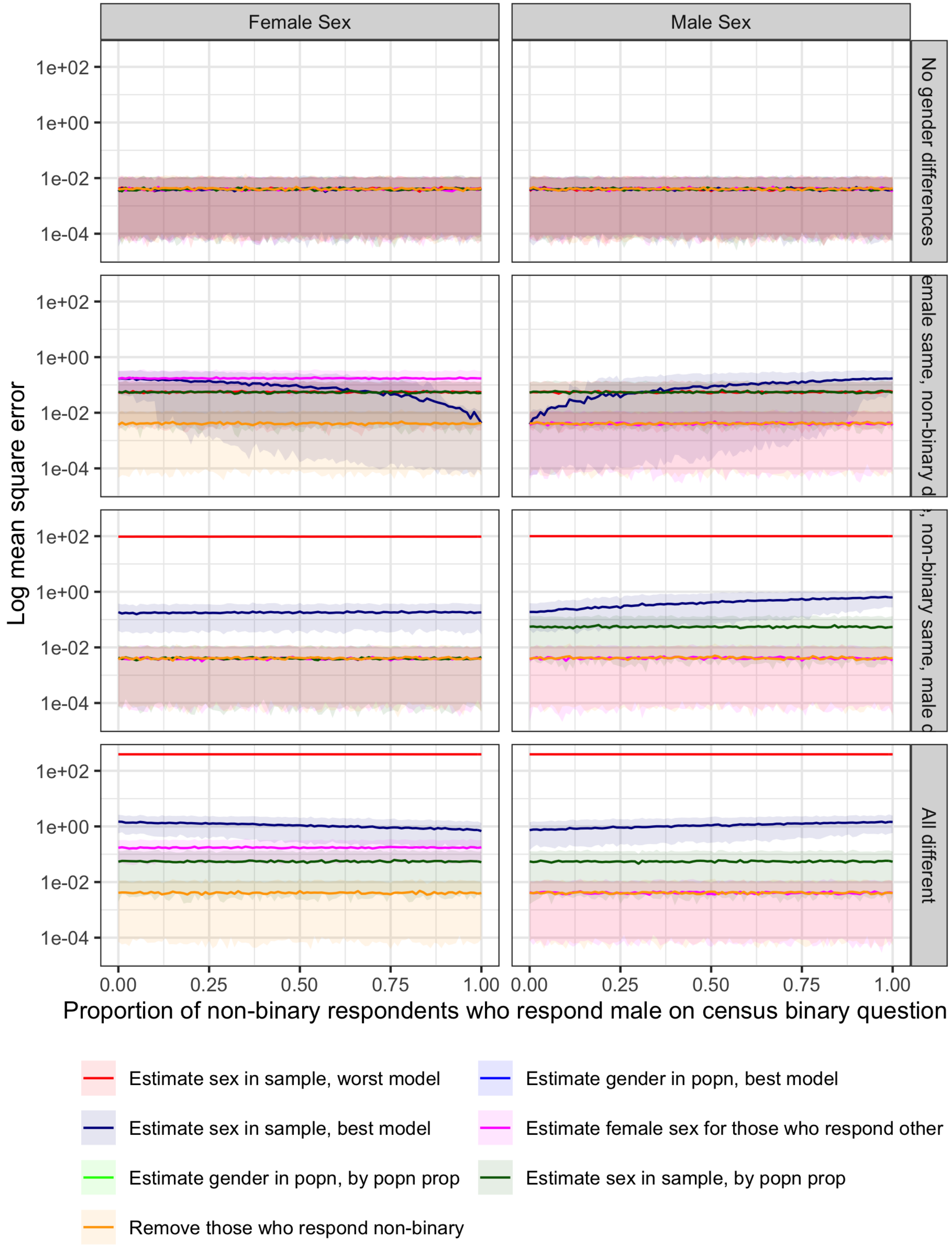

Figure 2: *Mean square difference between the true population estimate and the weighted estimate from the sample where the weights are created using the imputed gender variables. A lower mean square error is better. The colours represent different imputation strategies.*

a set of biological attributes, we would use imputed sex in waves of the survey that measure gender (but presumably these surveys would not have moved to measuring gender instead of sex). As official statistics organizations move to measuring gender, it seems likely that guidance will be provided on best use for the specific measurements they take, so we do not consider this further.

4 Looking forward

Measurement is an important aspect to scientific endeavours. It is a particular challenge to survey researchers and social scientists because the very constructs that we measure are changing in both importance and definition over time. This means that an appropriate measurement of a construct today might not be an appropriate measure of the construct 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 the same as sex assigned 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. What it does not do, however, is make broad recommendations for a one best way to measure sex or gender or a one best 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. Instead it is important to recognize that there is a compromise between ethical concerns, statistical concerns, and the most appropriate decision will be reflective of 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, we note that statistical and ethical concerns intersect, which implies that both facets need to be considered. 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. Our hope is that this is a useful resource to guide decision making for survey statisticians and survey administrators alike.

References

Maria Abascal. Us and them: Black-white relations in the wake of hispanic population growth. *American Sociological Review*, 80(4):789–813, 2015.

Alex A Ahmed. Trans competent interaction design: A qualitative study on voice, identity, and technology. *Interacting with Computers*, 30(1):53–71, 11 2017. ISSN 0953-5438. doi: 10.1093/iwc/iwx018. URL `https://doi.org/10.1093/iwc/iwx018`.

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

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

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

Jessica J Cameron and Danu Anthony Stinson. Gender (mis) measurement: Guidelines for respecting gender diversity in psychological research. *Social and Personality Psychology Compass*, 13(11):e12506, 2019.

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

Sophie Collyer, Maury Matthew, Lily Bushman-Copp, Irwin Garfinkel, Lauren Kennedy, Kathryn Neckerman, Julien Teitler, Jane Waldfoger, and Christopher Wimer. The State of Poverty and Disadvantages in New York City, 2020.

Columbia Population Research Center, 2012-2020. URL `https://cprc.columbia.edu/content/new-york-city-longitudinal-survey-wellbeing`.

Naomi Ellemers, Russell Spears, and Bertjan Doosje, editors. *Social Identity: Context, Commitment, Content*. Blackwell Publishers: Oxford, UK, 1999.

Naomi Ellemers, Russell Spears, and Bertjan Doosje. Self and social identity. *Annual Review of Psychology*, 53(1):161–186, 2002.

Office for National Statistics, 2020. URL `https://www.ons.gov.uk/census/censustransformationprogramme/questiondevelopment/sexandgenderidentityquestiondevelopmentforcensus2021`.

Andrew Gelman and Thomas C Little. Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23:127–135, 1997.

Howard M Glasser and John P Smith III. On the vague meaning of "gender" in education research: The problem, its sources, and recommendations for practice. *Educational Researcher*, 37(6):343–350, 2008.

Jessica Holzberg, Renee Ellis, Matthew Virgile, Dawn Nelson, Jennifer Edgar, Polly Phipps, and Robin Kaplan. Assessing the feasibility of asking about gender identity in the current population survey. results from focus groups with members of the transgender population. *Washington, DC: US Bureau of Labor Statistics. Available at: https://www. bls. gov/osmr/research-papers/2017/pdf/st170200.pdf (accessed September 2019)*, 2017.

Lauren Kennedy and Andrew Gelman. Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample. *Psychological Methods*, 2020.

Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018. doi: 10.1145/3274357. URL `https://doi.org/10.1145/3274357`.

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

Carolyn A Liebler, Sonya R Porter, Leticia E Fernandez, James M Noon, and Sharon R Ennis. America's churning races: Race and ethnicity response changes between census 2000 and the 2010 census. *Demography*, 54(1):259–284, 2017.

Esther L Meerwijk and Jae M Sevelius. Transgender population size in the united states: a meta-regression of population-based probability samples. *American Journal of Public Health*, 107:e1–e8, 2017.

Merriam Webster, 2020. URL `https://www.merriam-webster.com/dictionary/misgender`.

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

Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press, 2018.

Federal Interagency Working Group on Improving Measurement of Sexual Orientation and Gender Identity in Federal Surveys. Current measures of sexual orientation and gender identity in federal surveys. Technical report, 2016.

Ontario Human Rights Commission, 2000. URL `http://www.ohrc.on.ca/en/questions-and-answers-about-gender-identity-and-pronouns`.

David K Park, Andrew Gelman, and Joseph Bafumi. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, pages 375–385, 2004.

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

Aliya Saperstein, Andrew M Penner, and Jessica M Kizer. Making the most of multiple measures: Disentangling the effects of different dimensions of race in survey research. *American Behavioral Scientist*, 60:519–537, 2016.

Tom W Smith and Jaesok Son. Transgender and alternative gender measurement on the 2018 general social survey. 2019.

Statistics Canada. Sex at birth and gender: Technical report on changes for the 2021 Census. Technical report, 2020. URL `https://www12.statcan.gc.ca/census-recensement/2021/ref/98-20-0002/982000022020002-eng.cfm`.

United States Census Bureau. American community survey: Why we ask questions about sex, 2020. URL `https://www.census.gov/acs/www/about/why-we-ask-each-question/sex/`.

Nicholas Vargas and Kevin Stainback. Documenting contested racial identities among self-identified latina/os, asians, blacks, and whites. *American Behavioral Scientist*, 60(4): 442–464, 2016.

Stephen Voss, Andrew Gelman, and Gary King. The polls—a review: Preelection survey methodology: Details from eight polling organizations, 1988 and 1992. *Public Opinion Quarterly*, 59(1):98–132, 1995.