DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 54, ARTICLE 35, PAGES 1125–1158
PUBLISHED 5 JUNE 2026
https://www.demographic-research.org/Volumes/Vol54/35/
DOI: 10.4054/DemRes.2026.54.35

Research Article

Formatting the two-step gender measure: Experimental insights from the United States

Tessa Holtzman

Aliya Saperstein

Contents

Demographic Research: Volume 54, Article 35
Research Article

Formatting the two-step gender measure: Experimental insights from the United States

Tessa Holtzman[1]

Aliya Saperstein[2]

Abstract

BACKGROUND
Measuring sex and gender is central to demographic research and studies of social inequality. Best-practice guidelines for surveying English-speaking adults recommend a two-step approach, which uses two items to distinguish sex at birth from current identity. Additional research is needed to understand whether specific formatting decisions affect response distributions or measures of data quality.

OBJECTIVE
We explore three understudied formatting choices: (1) Should sex or gender terms be used for current identity responses, (2) should both items appear on the same page or different pages in online surveys, and (3) in which order should the response options be presented? By answering these questions, we provide practical guidance for researchers applying the two-step measure, and evidence on how robust the measure is to various formatting differences.

METHODS
We draw on three survey experiments conducted on two US adult samples: a sexual and gender minority (SGM) sample (n = 2,402) and a general population sample (n = 1,491).

CONCLUSIONS
Our results suggest researchers using the two-step approach can improve data quality by (1) using gender terms for current identity, and (2) using a page break in online surveys. Varying the response option order does not have measurable effects on data quality in our samples, which frees researchers in most survey contexts to consider nontraditional response orders that have other advantages.

[1] Stanford University, Stanford, California, USA. Email: tessamh@stanford.edu.
[2] Stanford University, Stanford, California, USA.

CONTRIBUTION
Some formatting changes can improve data quality, especially for SGM respondents, but small differences in how the two-step measure is formatted do not compromise comparability across studies or overall performance in general population surveys.

1. Introduction

Questions about gender are among the most common demographic items in surveys because gender is a primary categorical distinction that people use to interpret the social world and reproduce social inequality (Ridgeway 2011). Despite their ubiquity, the design of gender questions has long been critiqued not only because of conceptual conflation with sex but also because the gap between typical binary categorical measures and the diversity of gendered experience has become increasingly clear (Alexander, Bolzendahl, and Wängnerud 2021; Ridgeway and Saperstein 2024; Westbrook and Saperstein 2015). At the same time, sex and gender measurement has become a flash point for political contestation in the context of broader "anti-gender" movements (Korolczuk, Graff, and Kantola 2025). The most recent and high-profile example is US Executive Order 14168, issued in January 2025, which mandates that the federal government recognize only two unchangeable biological sexes: male and female. This shifting terrain between academic consensus and political directives presents challenges for individual researchers designing and conducting their own surveys. Although federal funding rules may constrain how researchers measure and report gender data in the United States, scientific research that recognizes both sex and gender diversity must continue.

We build on research among English-speaking adults that recommends a 'two-step' approach to gender measurement, which asks separately about sex assigned at birth and current gender identity (NASEM 2022). Although adoption of the two-step approach has been increasing around the world (Bornatici et al. 2025), its implementation varies both between and within countries. For example, in the United States, some federally sponsored studies have used gender terms when asking about current identity (e.g., the All of Us Survey and the 2018 General Social Survey [GSS]), while others have used sex terms (e.g., the US Census Bureau's Household Pulse Survey [HPS] and the National Crime and Victimization Survey [NCVS]). Other variations include whether the items are asked sequentially or on the same page, and how nominal response options are ordered (NASEM 2022). To our knowledge, there is no existing evidence that directly compares data quality across these key implementation considerations. As a result,

questions remain about how these design details influence reliability, inclusivity, and comparability across surveys with different formats.

In this study, we conduct survey experiments to address three fundamental design decisions for implementing the two-step measure and compare outcomes between two US national samples: a survey of lesbian, gay, bisexual, and transgender (LGBT) adults and a survey designed to be representative of the general adult population. We consider the following questions: First, should response options for the current gender identity question use sex-based terms (e.g., female/male) or gender-based terms (e.g., woman/man)? Second, in online surveys, should the two questions be presented on the same page or separated by a page break? Third, how should response categories be ordered within each question? By systematically evaluating these possibilities with respect to response distributions and data quality, we demonstrate that, although the two-step measure is generally robust to minor design differences, formatting choices do influence how respondents interpret and engage with these items. Our findings provide practical guidance for researchers seeking to collect high-quality, inclusive gender data in surveys.

2. Literature review

Sex and/or gender have long served as core measures for understanding population composition and inequality. In fact, sex has been included in the US Census since its inception in 1790 (Pao et al. 2025). Although sex and gender are conceptually distinct, they are often conflated (Cowan 2005; Hammarström and Annandale 2012; Wickes and Emmison 2007), especially among cisgender people whose gender identity does not differ from their sex assigned at birth (NASEM 2022). In contrast, gender scholars generally agree that sex and gender are distinct, that both encompass more than two categories, that self-identification may not match external classifications, and that both identities and classifications can shift over time (Westbrook and Saperstein 2015). Sex is generally understood to refer to divisions based on genitals, chromosomes, and/or hormone levels, and gender as behaviors stereotypically associated with a sex category, though gender is not determined by sex (West and Zimmerman 1987).

It is both theoretically and practically important to measure gender and sex as distinct concepts. Over the past 15 years, scholars and governments around the world have worked toward implementing more inclusive measurement practices, and a growing number of censuses and surveys now include dedicated sexual orientation and gender identity (SOGI) questions (Guyan 2022). Two recent United States National Academies of Sciences, Engineering, and Medicine (NASEM) consensus reports emphasize the importance of standardized SOGI data collection to both represent population diversity

and assess disparities in well-being (NASEM 2020, 2022). Although countries have employed different question formats, which can produce different population estimates (see, e.g., Biggs 2026 on such differences within the United Kingdom), there is broad agreement that the most inclusive and valid approach is some form of two-step measure, which asks separately about sex assigned at birth and current gender identity (Bornatici et al. 2025; Ridgeway and Saperstein 2024). Cross-tabulating these responses allows researchers to identify cisgender and transgender women and men, as well as respondents outside the gender binary. Studies in both gender minority-focused and general population samples, within and outside the United States, consistently find that a two-step approach yields more reliable and valid data than single-item sex or gender questions (see NASEM 2022 for a review).

Although the two-step approach is an improvement over past measurement practices, it is not a panacea. The two-step is favored over various one-step alternatives (e.g., just asking about current identity) because it maintains continuity with older binary sex data, facilitates skip patterns for sex-organ-specific questions (e.g., pregnancy), helps identify people with transgender experience who do not use the term 'transgender,' and aids in population weighting given the current absence of authoritative national benchmarks for gender identity and transgender status. However, the question about sex assigned at birth is a single, distant proxy for multiple relevant biological characteristics and, by definition, it does not allow for change over time, as occurs for many dimensions of biological sex, including secondary sex characteristics and hormone milieu (Ainsworth 2015). Further, concerns about data privacy and respect for respondent autonomy remain paramount, and asking about sex assigned at birth is not appropriate in all contexts (NASEM 2022). For example, asking about sex assigned at birth can be intrusive and unnecessary, especially in contexts where confidentiality cannot be guaranteed, such as the workplace or when applying for basic social services (Ridgeway and Saperstein 2024).[3] The two-step measure also does not capture all aspects of sex or gender that may be relevant to particular populations or research questions (e.g., awareness of intersex status or gender expression). Despite these limitations, the two-step measure remains one of the most validated ways of measuring gender in survey research.

Building on the evidence reviewed by expert consensus and presented in the NASEM (2022) report, our study provides guidance on specific formatting choices researchers must make when using the NASEM-recommended two-step approach. Although the NASEM (2022) report provides robust evidence in favor of the two-step approach overall, it identifies several areas where evidence on implementation is lacking. We begin to fill those gaps by examining (1) the wording of the answer options in the current-identity question, (2) whether to present sex-assigned-at-birth and current-

[3] One alternative in such contexts is to ask about current gender identity (e.g., man, woman, nonbinary) alone or alongside transgender status (yes/no, 'Do you have a transgender history or consider yourself transgender?').

identity questions on the same page or separate pages, and (3) the optimal response ordering for both items.

2.1 Current identity response terminology

The 2022 NASEM report recommends that researchers use sex terms (female/male) for both the sex-assigned-at-birth and current-identity items, though it notes there is no direct evidence to support this decision. Sex terms have been the most common approach in practice in the United States and abroad (NASEM 2022). This practice reflects an assumption that it is preferable to keep the response options parallel across the two-step items, particularly for cisgender respondents, who are the overwhelming majority of respondents in general population surveys and censuses.

Yet the use of sex terms may have data quality consequences, especially in studies of English-speaking adults. For the sex-assigned-at-birth item, sex categories match the terminology used on birth certificates in the United States (and elsewhere). For the current-identity item, however, gender terms may be more appropriate. Although many languages do not have separate words for sex and gender (Bornatici et al. 2025) and, even in those that do, many respondents may perceive little difference between sex-based and gender-based wording, the distinction may be particularly salient for transgender people. For all English-speaking participants, when asked about current gender and faced with response options such as "female," "male," "transgender," and "none of these," someone who identifies as a man or woman might choose "none of these" and be counted among nonbinary respondents, which would not accurately reflect their gendered sense of self.

Potential evidence of this pattern appears in a recent GSS longitudinal panel. In 2018, the GSS used gender terms ("woman/man"), while in 2020 it used sex terms ("female/male"). Among respondents in the GSS longitudinal panel who changed answers between waves, more than one-third shifted from "woman/man" to "none of these" (Saperstein 2022). Although some of these shifts could reflect the contemporaneous rise in nonbinary identification, the pattern is striking enough to raise concerns about the validity and reliability of measures when surveys switch between sex- and gender-based terminology in gender identity questions. Given existing evidence, we might expect that using gender terms for the identity question will result in better data quality, especially in an LGBT-focused sample.

2.2 Pagination of the two-step measure

Another unresolved design feature of the two-step measure is whether both items, sex assigned at birth and current identity, should appear on the same survey page, particularly in online surveys. For general population surveys of English-speaking adults, the 2022 NASEM report recommends asking sex assigned at birth before gender identity, following a chronological order from past to present, but does not provide guidance on whether the items should be presented together on the same page or separately.

To our knowledge, only one study has speculated about the influence of page placement on responses. In cognitive interviews, Bauer et al. (2017) find that cisgender participants generally do not perceive a distinction between the two parts of the two-step measure, while transgender participants do. The authors conclude that transgender respondents' understanding may depend on their being able to view both items at once (though the study does not vary item placement, so this conclusion is not empirically tested). When the two-step items appear on the same page, transgender respondents can see they are being asked about both birth sex and current identity. This may reduce the likelihood that they enter their gender identity in response to the sex-at-birth item. Cisgender respondents may also find the same-page format clearer, since it makes it obvious the survey is not asking the same question twice.

At the same time, many large surveys (e.g., the GSS) present, and survey best practices recommend, placing one item per page. This recommendation assumes that single item per page formats reduce respondent fatigue, improve the respondent survey experience, and reduce respondent speeding, though explicit evidence on pagination is mixed (Manfreda, Batagelj, and Vehovar 2006; Toepoel, Das, and Van Soest 2009). Further, whether this convention affects data quality for sex and gender items specifically has not been tested. This leads to competing expectations of whether presenting the two items on the same page or separate pages will lead to fewer data quality issues for cisgender and transgender participants alike.

2.3 Ordering of response categories

Little empirical research has examined how the ordering of nominal response categories affects demographic identity items. Yet this seemingly minor design choice has the potential to reinforce existing hierarchies and introduce measurement error. Primacy effects in survey response lists are well documented (Galesic et al. 2008), but the solution, randomization, is rarely applied to demographic identity items. Instead, surveys rely on conventions like alphabetical ordering (e.g., Wells, Bailey, and Link 2013) or ordering by expected population size, as in the US Census Bureau race items (NASEM 2022).

Despite these conventions, the 2022 NASEM report notes that many surveys list male first and female second when asking about sex assigned at birth – a practice that is neither alphabetical nor justified by relative population size. Category ordering in surveys and their research products can carry social meaning (Loveman 2014), in this case reinforcing systems of patriarchy. In response, and because it is consistent with both alphabetical and expected population-based reasoning, NASEM recommends listing female first.

Some surveys have adopted this approach. For example, in the United States, the All of Us survey lists female before male (Cronin et al. 2019), and the 2018 GSS did the same based on a survey experiment that randomly varied the answer order in an online sample of US adults (Saperstein and Westbrook 2019). England and Wales also listed female first in its 2021 Census, after critiques of the male-first default order were raised in preparatory focus group research (Office of National Statistics 2021). Given the relative lack of rigorous testing between competing approaches, we cannot offer any a priori expectations about optimal response option ordering before assessing the effects empirically.

3. Data and methods

The data for this study come from two US samples: (1) a sample designed to be representative of LGBT-identified adults, and (2) a sample designed to be representative of the general adult population. The LGBT sample is the primary focus of our analysis, as it contains greater variation in gender identities and therefore provides greater leverage for detecting differences in response patterns across experimental conditions. At the same time, the general population sample serves as an important robustness check for how these items perform among primarily cisgender respondents.

Experimental surveys were fielded in May 2023 to both samples with identical question order, wording, and answer options. The LGBT sample (n = 2,402) was sponsored through the Time-Sharing Experiments in the Social Sciences (TESS) program and recruited participants who had previously identified as LGBT. Within this sample, 44% of responses were drawn from the AmeriSpeak adult panel (n = 1,056), and 56% from the Lucid panel (n = 1,346). The general population sample (n = 1,491) was recruited via Prolific using quota sampling by age group, racial identification, and binary sex category. Table 1 provides descriptive statistics on the demographic variables discussed in the main text for both samples. Appendix Table A-1 provides descriptive statistics across the complete set of demographic variables explored in all our analyses.

Table 1: Demographic characteristics for US LGBT and adult samples

	US LGBT Sample	*US Adult Sample*
N	2,402	1,491
Sex assigned at birth		
Female	1,450 (60.4%)	770 (51.6%)
Male	952 (39.6%)	721 (48.4%)
Gender		
Woman	1,301 (54.2%)	749 (50.2%)
Man	875 (36.4%)	714 (47.9%)
Transgender	106 (4.4%)	13 (0.9%)
Different term	120 (5.0%)	15 (1.0%)
Region		
South	914 (38.1%)	577 (38.8%)
West	558 (23.2%)	334 (22.5%)
Midwest	547 (22.8%)	314 (21.1%)
Northeast	383 (15.9%)	262 (17.6%)
Missing	0 (0.0%)	4 (0.3%)
Sexual orientation		
Straight	131 (5.5%)	1,268 (85.0%)
Bisexual	1,298 (54.0%)	128 (8.6%)
Lesbian or gay	794 (33.1%)	55 (3.7%)
Different term	179 (7.5%)	40 (2.7%)
Race/origin		
White or European American	1,801 (75.0%)	1,165 (78.1%)
Hispanic, Latino, or Spanish origin	177 (7.4%)	82 (5.5%)
Black or African American	451 (18.8%)	192 (12.9%)
Asian or Asian American	91 (3.8%)	108 (7.2%)
American Indian or Alaska Native	99 (4.1%)	23 (1.5%)
Middle Eastern or North African origin	23 (1.0%)	8 (0.5%)
Native Hawaiian or Pacific Islander	22 (0.9%)	3 (0.2%)
Another race/origin not listed	20 (0.8%)	4 (0.3%)
Age		
18–29	853 (35.5%)	328 (22.0%)
30–44	857 (35.7%)	408 (27.4%)
45–59	364 (15.2%)	415 (27.8%)
60+	328 (13.7%)	340 (22.8%)
Employment status		
Employed	1,561 (65.0%)	745 (69.4%)
Not in labor force	548 (22.8%)	221 (20.6%)
Unemployed (looking for work)	293 (12.2%)	107 (10.0%)
Missing	0 (0.0%)	418 (28.03%)

Notes: Nominal variables in Table 1 and all following figures and tables are ordered from highest population prevalence in the United States to lowest population prevalence. For the general adult sample, we derived region from reported three-digit zip code. The four respondents who are coded as missing region provided an uncodable zip code. Racial categories are not mutually exclusive and so percentages may not sum to 100.

As expected, given what is known about the LGBT population in the United States, our LGBT sample predominately identifies as women (54%) and as bisexual (54%) and is relatively young (71% below age 45). In contrast, the general adult sample overwhelmingly identifies as straight (85%) and is distributed relatively evenly by binary sex and age, per sample quotas (see Table 1). The majority of respondents in both samples identified their race or origin as "White or European American" (75% in the LGBT sample and 78% in the general adult sample).

Both surveys included the two-step measure following the NASEM (2022) recommendations. The sex-assigned-at-birth question offered only "female" and "male" as response options, as is recorded on birth certificates in the United States. The current-identity item response options differed by condition, as described below, but always included the category "transgender" and a write-in option prefaced by "I use a different term." Importantly, this residual category avoids dehumanizing labels common in survey items, such as "none of these," "something else," or "other."

The current-identity item does restrict respondents to selecting a single response, which creates a forced choice for respondents who identify with more than one gender category. For example, among respondents who identify as both transgender and women, some may choose to identify their current gender as "transgender," while others may choose to identify as "woman" or to write in an answer with both terms. This aspect of the NASEM-recommended two-step has been critiqued (e.g., Restar et al. 2024), and many argue "transgender" is better seen as a status, experience or modality rather than a gender category (see, e.g., Lindqvist, Sendén, and Renström 2021). However, in the United States, it remains common to include "transgender" as a stand-alone option for people who see the term "transgender" as central to their gender identity. As our data show, a substantial number of US respondents continue to choose "transgender" even when offered the opportunity to write in a preferred term (or terms). Out of 2,402 people in the LGBT sample, 106 chose to identify as transgender while 34 identified with a binary gender that differed from their sex assigned at birth. Further, out of the 120 respondents who opted to provide a different term, just two provided a combined term like "trans woman" or "trans man." There also is no evidence that a forced-choice design for current identity hinders a full count of the transgender population (NASEM 2022). This is because the two-step measure allows for the cross-tabulation of responses between current gender and sex at birth such that researchers can inclusively identify transgender respondents who identify explicitly as "transgender" and transgender respondents who identify with a term or category other than "transgender" as their current identity. Thus, we retained the single-response format of the current-identity item and focused our experiments on issues identified in the NASEM (2022) report as areas where there was little to no existing evidence on which to base formatting decisions.

3.1 Experimental design

Three experiments were embedded in the surveys, each using a between-subjects design with random assignment.

Experiment 1. In the first experiment, participants were randomly assigned to one of two conditions: a condition where they saw sex terms (female/male) or a condition where they saw gender terms (woman/man) in the answer options for current identity (see Figure 1). The current-identity question was otherwise identically worded, and the sex-assigned-at-birth question remained unchanged.

Figure 1: Response wording for the term type experiment

Gender terms	Sex terms
1. What sex were you assigned at birth, on your original birth certificate? *Female* *Male*	1. What sex were you assigned at birth, on your original birth certificate? *Female* *Male*
2. What is your current gender? *Man* *Transgender* *Woman* *I use a different term (please specify)*_____	2. What is your current gender? *Female* *Male* *Transgender* *I use a different term (please specify)*_____

Notes: Experiment 1 varied whether sex or gender terms were used for the second item in the two-step question. Response options are shown alphabetically, representing one of the four conditions in the response order experiment (Experiment 3). See Appendix Table A-2 for the full set of conditions used in the response order experiment.

Experiment 2. In the second experiment, participants were randomly assigned to either the same-page condition or the different-page condition. In the different-page condition, the sex-assigned-at-birth question was shown first, followed by the current-identity question on a new page. In the same-page condition, the two questions were shown sequentially on the same page. No other questions appeared on the page in either condition. A confirmation question always followed the two-step measure on a separate page (i.e., either a second or third page), allowing respondents to correct any inadvertent responses.[4]

Experiment 3. The third experiment had four between-subject conditions. Each condition contained a different order pattern for the response options for both the sex-assigned-at-birth and current-identity items. The four ordering conditions were (1) alphabetical order, (2) population prevalence from high to low, (3) population prevalence

[4] The 2022 NASEM report provides guidance that the two-step measure should be followed by a confirmation question, which is typically asked only of noncisgender responses because of concerns about errors by cisgender people artificially inflating transgender population estimates. We ask this confirmation question of all participants, in part because we use the number of corrections as a measure of data quality.

from low to high, and (4) random (see Appendix Table A-2 for the specific response ordering in each of these conditions). The residual category ("I use a different term") appeared last across all four conditions, and there was not within-subject variation (e.g., participants saw the same ordering across both two-step items).[5]

3.2 Outcome variables

For each experiment, we analyze two outcomes: (1) the distribution of gender identities across conditions, and (2) the error or correction rates observed based on the follow-up confirmation item. Gender identity is collected using four response categories: "woman/female," "man/male," "transgender," and a "different term." In the results below, we examine the full four-category distribution to understand the potential for response differences across experimental conditions. Second, we focus on the prevalence of write-ins, using a binary indicator equal to one if a respondent selected "I use a different term," and zero otherwise.

Write-ins provide important information about response option fit: Fewer write-ins suggest that the offered categories better capture respondents' identities. Write-ins also are less desirable because they require manual coding, which can introduce measurement error by substituting researcher decision-making for respondent responses. A write-in option is not recommended for the sex-at-birth question, so this measure applies only to the current-identity question. As expected, gender identity write-ins were more common overall in the LGBT sample (5%) than in the general population sample (1%). In both samples, the most common write-in among respondents who preferred a different gender term was "nonbinary" or a close variant, a point we return to later in the discussion.

Error or correction rates for the two-step measure were assessed using a follow-up confirmation question. In this study, every respondent received such a confirmation:

Just to confirm, you said you were assigned [X] *at birth and you currently identify as* [Y], *is that correct?*

Response options included "yes," "birth sex incorrect," "current gender incorrect," and "both incorrect." Respondents who selected any "incorrect" option were given the opportunity to re-answer the relevant item(s). Respondents who chose not to answer either or both items were asked to confirm that they did not wish to share.

[5] The response order was also tested for questions about sexual orientation and racial identity. These results, and the results from a fourth experiment, which varied the order for scales measuring health and femininity and masculinity, are explored elsewhere (Saperstein and Ren 2026; Saperstein and Sobotka 2025).

We coded any respondent selecting an "incorrect" option as having requested a correction. Within this group, a true correction was recorded if the revised answer differed from the initial one (e.g., a respondent who was assigned female at birth reported identifying as "man" but then changed this response to "woman"). In contrast, if a respondent indicated a correction was needed but did not change their initial response (e.g., selecting "man" both times), this was coded as a false correction.[6] Although our data do not reveal why respondents made false corrections, these types of corrections were common. Overall, corrections were rare but more frequent in the LGBT sample (n = 44, 1.83%) than in the general adult sample (n = 8, 0.54%). More than half (59%) of the respondents who requested corrections in the LGBT sample did not change their response for at least one of the items they reported as incorrect, nor did 38% of the general population sample.

We focus on true corrections in the text because we believe this to be the most theoretically relevant measure, but our conclusions are not sensitive to this approach, and we include distributions for all types of corrections in the tables below and in the Appendix. We also consider differences in the distribution of corrections across conditions for participants who initially reported noncisgender identities, given that standard practice is to ask confirmation questions only of these respondents. In our data, 75% of the corrections in the general adult sample followed the expected pattern of changing from noncisgender identities to cisgender identities, compared to just 11% of corrections in the LGBT sample. We return to address the implications of these results for the use of confirmation questions in our discussion.

3.3 Analytic sample

Data for the LGBT sample were collected and cleaned by NORC at the University of Chicago. As part of preparing the public data file,[7] NORC removed 114 respondents who consented to participate and were exposed to at least one of the experimental manipulations presented in this paper because either (a) they were flagged for low data quality (speeding and/or high missingness), or (b) they did not click through to the end of the survey instrument (NORC 2023). We chose to exclude one additional case from the public-use data because the respondent declined to share their gender identity and sex assigned at birth. This individual was also missing responses to other demographic variables and completed the survey in less than one minute.

[6] If a respondent reported both items were incorrect, the respondent could be categorized as having made both a true correction and a false correction if they changed their initial response for one of the items but not for the other.

[7] File available at https://tessexperiments.org/study/saperstein1530.

We examine the distribution of these 115 cases by our experimental conditions because their exclusion could bias the estimation of treatment effects (see Appendix Table A-12). We find no association between case exclusion and our conditions for Experiment 1 (sex versus gender terms, $p = 0.853$) or Experiment 3 (response option order, $p = 0.427$). We observe a small association with the page-break experiment: 5.6% of respondents in the different-page condition were excluded, compared to 3.4% in the same-page condition ($p = 0.006$). This pattern is not associated with balance on demographic characteristics across conditions (Appendix Table A-13), and it does not meet the conventional threshold for statistical significance when the speeding/missing data cases ($n = 95$) are separated from the drop outs ($n = 15$),[8] suggesting that the observed association with the page-break condition is not connected to a particular type of respondent and may have arisen by chance. Nevertheless, to mitigate concerns, we conduct robustness checks for the page-break analyses using an expanded sample that includes all 115 excluded cases (see Appendix Table A-14). Overall, we find that including the 115 excluded cases in the page-break analyses does not meaningfully change the page-break results or our conclusions.

For the general adult population sample, the data were minimally cleaned to remove duplicate responses ($n = 10$). We also drop respondents who did not complete the survey or returned it without seeking payment ($n = 83$) since these responses could be interpreted as a revocation of consent. We find no evidence of bias introduced by excluding these cases (see Appendix Table A-15).

3.4 Analytic approach

To evaluate differences across the experimental conditions, we use chi-square tests of independence for each of our outcome measures. The null hypothesis for each test reflects an expectation of no difference between our experimental conditions. In our results, we report the continuous *p*-values derived from these tests (rather than relying on discrete conventional significance thresholds), together with the test statistic, degrees of freedom, and sample size. We provide this information in addition to the *p*-value so that readers can consult the chi-square distribution directly and assess the uncertainty around our estimates.[9]

[8] Though 20 cases were dropped for incompleteness, 5 of these cases completed our survey questions but dropped out before completing post-treatment measures of political ideology and religion required by TESS. For this analysis, we count these 5 cases as complete.

[9] The output produced by the code provided in the replication package additionally shows the expected and observed counts for each chi-square test. We used StataSE version 19.5 for all statistical analysis.

For analyses where the outcome is true corrections, and the associated chi-square *p*-value exceeds 0.05 but is less than 0.50, we conduct post hoc power analyses to assess the sample size required to detect a measurable effect with adequate power. We present these results to contextualize the magnitude of the observed differences and indicate how large a sample would need to be before the differences across conditions we observe in our sample could produce results that are statistically distinguishable from zero. Power calculations are conducted using G*Power, assuming $\alpha = 0.05$ and power = 0.80, and incorporating the degrees of freedom and effect size derived from our data, with effect size expressed as Cohen's ω, which is calculated using the following formula:

$$\omega = \sqrt{\chi^2/_n}.$$

Given the experimental nature of this paper, when a result has a *p*-value below the conventional threshold of $p < 0.05$, we interpret this to mean the outcome differences are statistically distinguishable from zero in our samples. For the sex/gender terms experiment (Experiment 1), we also provide average marginal effects from bivariate logit models with 95 percent confidence intervals to help readers gauge the uncertainty of our estimates and to provide an interpretation of magnitude. We do not run the same logit analysis for Experiment 2 because of small cell sizes for our main outcome of interest, true corrections. We do not run the logit analysis for Experiment 3 because the response order experiment is not a 2 × 2 comparison.

As a robustness check, we replicate the chi-square tests within demographic subgroups, defined by sex at birth, age, sexual orientation, race/origin, region, education, income, marital status, employment status, political ideology, and religion.[10] These analyses allow us to examine whether patterns observed in the full sample hold across key subpopulations and to evaluate the generalizability of the findings. Missing data on these characteristics are handled using pairwise deletion.

Several of our outcomes are relatively rare events, especially when subset by experimental condition. To complement the chi-square results, we conduct robustness checks using Fisher's exact tests for all correction analyses and, for other analyses, when any expected cell sizes are less than five.[11] These supplementary analyses are shown in

[10] Standard profile data from Prolific does not include education, income, marital status, political ideology, or religion, so those comparisons are limited to the LGBT sample.

[11] For analyses using logistic regression, when any cell contains fewer than 10 cases, we estimate Firth logistic regression models in addition to logistic regression models. Firth regressions reduce issues of small sample bias in maximum likelihood estimation models and are appropriate when analyzing rare events (Firth 1993). These supplementary models suggest that our conclusions are not sensitive to the assumptions of logistic regression. The results from the Firth regressions are presented in Appendix Table A-4.

the Appendix tables and suggest our conclusions are not sensitive to the assumptions of a specific test statistic.

4. Results

We begin by presenting results from Experiment 1, comparing the full distribution of current-identity responses, the prevalence of write-ins, and corrections between the sex terms and gender terms conditions to assess whether the wording difference affects responses. We then turn to Experiment 2, which tests whether asking the two-step items on the same page improves clarity compared to separating the items. Finally, Experiment 3 evaluates the consequences of different response option ordering.

4.1 Identity terminology: Sex versus gender

Across the outcome measures, our expectations for Experiment 1 were generally supported (Table 2). Although the four-category distribution of current-identity responses did not differ statistically by condition among the LGBT sample (X^2 (3, n = 2,402) = 5.0, p = 0.17), we observed more write-ins in the sex terms conditions (6%) than in the gender terms condition (4%) (X^2 (1, n = 2,402) = 4.7, p = 0.03). A similar pattern emerged for true corrections. Respondents in the LGBT sample made 11 such corrections: 82% in the sex terms condition compared to just 18% in the gender terms condition (X^2 (1, n = 2,402) = 4.9, p = 0.03).

It is also relevant to limit the corrections analyzed to participants who initially selected noncisgender responses because the opportunity to make a correction is typically offered only to survey respondents who initially answer the two-step measure in a way that is consistent with a noncisgender identity (NASEM 2022). When the corrections are limited in this way, the pattern above is mirrored: Of the eight true corrections among these respondents, 88% were in the sex terms condition and 12% were in the gender terms condition (X^2 (1, n = 260) = 4.1, p = 0.04). Given small cell counts, we interpret these results cautiously; however, all the corrections evidence for the LGBT sample points in the same direction: There are fewer corrections when respondents see gender terms instead of sex terms for the identity question.

Table 2: Main effects of identity response terminology (Experiment 1), US LGBT sample

	Outcome measure	Gender terms	Sex terms	*p*-value
	Gender identity distribution (%)			
4-category identity distribution	woman/female	55%	53%	
	man/male	36%	36%	
	transgender	5%	4%	
	different term/write-in	4%	6%	0.17
2-category identity distribution	all other categories (woman, man, transgender)	96%	94%	
	different term/write-in	4%	6%	0.03
	Total (*n*)	1,233	1,169	
	Correction distribution (n, ref cat = no corrections)			
Current identity corrections	all corrections	7	13	0.14
	true corrections	2	9	0.03 (0.03)
	false corrections	5	4	0.80 (1.00)
	Total	1,233	1,169	
Current identity corrections for noncisgender participants only	all corrections	3	8	0.17 (0.22)
	true corrections	1	7	0.04 (0.07)
	false corrections	2	1	0.51 (0.61)
	Total	124	136	

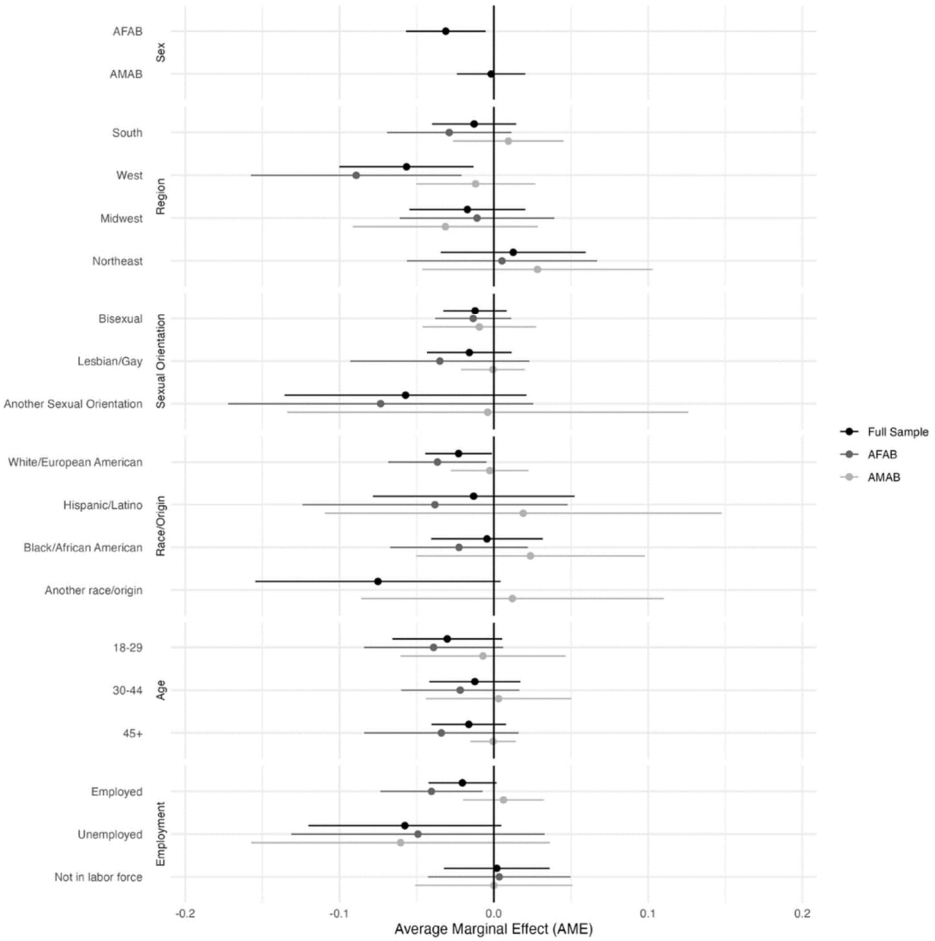
Notes: The *p*-values for chi-square tests are shown for all comparisons. The *p*-values from Fisher's exact tests are shown in parentheses for all analyses where at least one expected cell count falls below five.

Considering generalizability, these differences in write-in and correction frequency appear to be driven by respondents who are assigned female at birth (AFAB). Nearly 8% of AFAB respondents chose to write in a current identity in the sex terms condition, compared to 5% in the gender terms condition (X^2 (1, n = 1,450) = 5.9, p = 0.02).[12] In contrast, respondents assigned male at birth (AMAB) were equally likely to select the write-in option regardless of condition (X^2 (1, n = 952) = .02, p = 0.88). Figure 2 displays the average marginal effects of using gender terms (as compared to sex terms) for the LGBT sample, from separate bivariate logit models for each demographic characteristic shown in Table 1. Although we do not expect systematic differences across demographic subgroups, we include these analyses to assess the possibility that our full sample tests could be masking relevant subgroup variation. For example, our main effect could be the result of averaging two strong but opposing subgroup effects either within a given factor on its own (e.g., living in the West versus the South), or when we split the sample by sex

[12] We do not conduct the same analysis looking at true corrections because there are numerically too few true corrections in the sample to be powered to do so, but descriptively we see a similar pattern as the subgroup analysis for write-ins. For example, among assigned female at birth respondents, there were five true corrections to current identity made in the sex terms condition and none in the gender terms conditions. This pattern also holds descriptively for those assigned male at birth: Four true corrections to current identity were made in the sex terms condition and two were made in the gender terms condition.

assigned at birth. The full set of subgroup estimates – including all demographic variables – is reported in Appendix Tables A-3 and A-4.

Figure 2: Average marginal effect of using gender terms, US LGBT sample



Notes: Figure was produced using R Version 4.4.2. Estimates from bivariate logit models, with 95% confidence intervals. The average marginal effect should be interpreted as the percent point change in the frequency of write-in responses associated with using gender terms (as compared to sex terms), averaged across the subgroup. Estimates are shown for the full sample and separately by sex at birth for demographic characteristics available in both samples. Some demographic categories (e.g., race/origin and age) were collapsed to due to small cell sizes. The full set of subgroup estimates is reported in Appendix Table A-4. AFAB: assigned female at birth, AMAB: assigned male at birth.

The overall pattern in Figure 2 is clear: Using gender terms in the identity question yields fewer write-in responses across most demographic subgroups. To the extent there is contrary evidence, it is restricted to respondents who were assigned male at birth and has wide uncertainty.[13] For example, in the full sample, respondents who identified as Black or African American are equally likely to provide write-ins in either condition, but this averages response patterns in different directions by sex at birth: Black AFAB respondents are less likely to choose a write-in response with gender terms and Black AMAB respondents are more likely to do so, though the confidence intervals for both estimates overlap with zero. We see a similar contrast by sex at birth for employed respondents and respondents living in the South. However, even with these minor variations, we find no evidence that using sex terms offers a meaningful improvement in data quality for any subgroup in the LGBT sample.

Turning to the general population sample, no significant differences emerged across the outcomes (see Appendix Table A-5). Neither the four-category distribution of identity responses (X^2 (3, n = 1,491) = 4.2, p = 0.24), the prevalence of write-ins (X^2 (1, n = 1,491) = 1.7, p = 0.20), nor the number of true corrections (X^2 (1, n = 1,491) = 1.0, p = 0.32) varied by condition, though descriptively there were more write-ins and true corrections in the gender terms condition than the sex terms condition. A power calculation suggests that for general surveys of US adults with more than 12,000 participants, researchers should conduct additional testing to replicate our results in larger samples.[14] In smaller surveys, patterns of this magnitude cannot be distinguished from random variation and an identity question with gender terms performs just as well as one with sex terms even among non-LGBT respondents.

4.2 Pagination

We next examine whether the two-step items perform better when presented on the same page or across separate pages. Previous research with cisgender and transgender respondents (Bauer et al. 2017) speculates that misunderstanding and thus misclassification may be more likely when questions are split by a page break. However, we find no support for the same-page format. In addition to using the distribution and

[13] Cisgender bisexual women comprise the majority of the LGBT sample, making tests of subgroup differences among participants assigned male at birth relatively underpowered.

[14] There were 5 write-ins in the sex terms condition and 10 in the gender terms condition. There were no true corrections in the sex terms condition and 1 true correction in the gender terms condition. For both results, the p-values were large (0.20 and 0.32, respectively), and observed effects were small ($\omega \approx 0.034$ for write-ins and $\omega \approx 0.026$ for true corrections). Power calculations indicate that detecting an effect on the order of the write-in difference would require a sample of about 7,000 and an effect on the order of the true correction difference would require a sample of about 12,000.

data quality comparisons we introduced above (reported in Table 3), we also consider several possible explanations for the pattern of results. Though our follow-up analyses are only suggestive, the evidence tilts toward a different-page design.

Table 3: Main effects of pagination (Experiment 2), US LGBT sample

Outcome measure		*Same page*	*Different page*	*p-value*
Gender identity distribution (%)				
4-category identity distribution	woman/female	53%	56%	0.36
	man/male	37%	36%	
	transgender	5%	4%	
	different term/write-in	6%	4%	
2-category identity distribution	all other categories (woman, man, transgender)	95%	96%	0.21
	different term/write-in	6%	4%	
	Total (n)	1,208	1,194	
Correction distribution (n, ref cat = no corrections)				
Corrections to either sex at birth or current identity	all corrections	25	19	0.38
	true corrections	14	6	0.08
	false corrections	13	13	0.98
	Total	1,208	1,194	
Corrections to either sex at birth or current identity for noncisgender participants only	all corrections	15	7	0.22 (0.27)
	true corrections	12	3	0.05 (0.06)
	false corrections	5	4	0.99 (1.00)
	Total	145	115	

Notes: The *p*-values for chi-square tests are shown for all comparisons. The *p*-values from Fisher's exact tests are shown in parentheses for all analyses where at least one expected cell count falls below five.

Among the LGBT sample, the distribution of gender identity responses did not differ by condition (X^2 (3, n = 2,402) = 3.2, p = 0.36), nor did the prevalence of write-ins (X^2 (1, n = 2,402) = 1.6, p = 0.21). True corrections, however, were marginally more common in the same-page condition than in the different-page condition (70% versus 30%) (X^2 (1, n = 2,402) = 3.1, p = 0.08).[15] The overall effect of page formatting on data quality appears modest, but separate pages may help minimize corrections.[16]

[15] When limiting the true corrections analyzed to respondents who gave noncisgender answers initially, there were also more corrections in the same-page condition (8.3%) and fewer (2.6%) in the different-page condition (p = 0.05). This finding holds directionally when including the 115 cases excluded from the main analysis (8.2% corrections in the same-page condition and 4% in different-page condition), but the p-value increases to 0.15 (0.21 with the Fisher's correction).

[16] To detect an effect of the observed magnitude ($\omega \approx 0.036$) for true corrections would require a sample size of about 6,000, though an effect of that size is generally considered negligible. This is the case for all the true corrections analyses we explore in this section: All are of similar magnitude and would require sample sizes between 4,000 and 13,000 to observe effects that differ measurably from zero.

Exploring potential subgroup differences in the number of corrections yields no evidence of distinct response patterns.[17] For these comparisons, we use previously reported sex at birth from respondents' demographic profiles to compare results because reported sex at birth may have been affected by the experimental manipulation. Among previously identified AFAB respondents, there were nine true corrections (1.3%) in the same-page condition and three (0.4%) in the different-page condition (X^2 (1, n = 1,449) = 3.1, p = 0.08). Among previously identified AMAB respondents, the true corrections were five (1.0%) and three (0.7%), respectively (X^2 (1, n = 953) = 0.4, p = 0.53). Additional robustness checks by other demographic subgroups showed no major deviations from the overall trend (see Appendix Table A-6).

We also tested for interactions between the page-format experiment and the other experimental manipulations. For example, using sex versus gender terms in the current-identity item (Experiment 1) did not affect our conclusions about the page-break experiment. True corrections were marginally more frequent in the same-page condition (six corrections compared to one) when gender terms were used (X^2 (1, n = 1,233) = 3.5, p = 0.06). When sex terms were used, true corrections were also more common (eight compared to five) in the same-page condition (X^2 (1, n = 1,169) = 0.7, p = 0.41). Examining page format in combination with the response-ordering experiment (Experiment 3) likewise revealed no interaction (see Appendix Table A-7), as true corrections were descriptively more common in the same-page condition for all four response order conditions. We consider two additional possibilities for interactions between the page-break experiment and the response ordering. First, we examine if the page-break results were driven by respondents for whom the response option order for the sex and gender questions did not share corresponding first options.[18] For example, participants in the low-to-high prevalence order condition would have seen "male" listed first for the sex-assigned-at-birth item, and "transgender" listed first for the current-identity item. Second, we consider whether previously identified AMAB respondents who saw female/woman listed first in either item might be especially prone to mis-clicks in the different-page condition.[19] Neither of these possibilities explain our results (see Appendix Table A-7). For example, among AMAB respondents who saw "female" listed

[17] We focus on subgroup differences in the prevalence of true corrections because it had the smallest p-value among our three outcomes of interest. Given the numerical rarity of true corrections, these results should be interpreted as suggestive.

[18] 1,101 participants from the LGBT analytic sample saw response order options across the two items that did not share corresponding first options.

[19] 599 previously identified AMAB respondents from the LGBT sample saw female/woman first for either the sex-assigned-at-birth item, current-identity item, or both. See Appendix Table A-2 for examples of the conditions under which this occurred. In the random condition, whether female/woman appeared first in the current-identity item did not depend on where female appeared in the sex-at-birth item (and vice versa). There were participants for whom "female" appeared first for the sex at birth, and "man/male" appeared first for current identity.

first for either item, there were two corrections in the same-page condition and two corrections in the different-page condition (X^2 (1, n = 599) = 0.004, p = 0.95). Thus, we find no evidence that our page-formatting conclusion differs depending on other simultaneous formatting decisions.

We also explore survey speeding as a plausible explanation for the results, given prior work that finds survey pagination is related to completion time. This work suggests that respondents may rush more when multiple items appear on the same page. Page breaks may slow respondents down, reducing mis-clicks. Although we lack item-level timing data, total survey duration offers some suggestive evidence: Participants in the different-page condition took 10 to14 seconds longer on average than those in the same-page condition, depending on whether we consider all cases, including those removed in part because of speeding (t(2,512) = 1.04, 95% CI: –0.16 – 0.54, p = 0.30), or only our analytic sample (t(2,400) = –1.2, 95% CI: –0.59 – 0.13, p = 0.21). Among respondents who made corrections, this difference is more pronounced – those in the same-page condition completed the survey about 1.4 to more than 2 minutes faster than those in the different-page condition, comparing among all cases (t(52) = 0.68, 95% CI: –2.7 – 5.4, p = 0.50) and among our analytic sample (t(41) = –0.9, 95% CI: –6.9 – 2.5, p = 0.35), respectively.[20] These differences do not meet a conventional 0.05 p-value benchmark, but we wouldn't necessarily expect them to. Any speeding that occurred in the absence of a page break between the two-step measure would be minimal relative to the overall length of the survey, which did not vary by page-break condition.

Turning to the general population sample, the results are even clearer (see Appendix Table A-8). The four-category distribution of identity responses (X^2 (3, n = 1,491) = 1.3, p = 0.73) and prevalence of write-ins (X^2 (1, n = 1,491) = 0.64, p = 0.42) did not differ by conditions, but every correction (both true and false) occurred in the same-page condition (X^2 (1, n = 1,491) = 8.0, p = 0.01). Six of the eight (75%) corrections came from previously identified AFAB respondents, reducing concerns that the results are driven by AMAB respondents' mis-clicking. Consistent with the LGBT sample results, participants in the different-page condition took slightly longer to complete the survey (13 seconds on average) (t(1,494) = –1.3, 95% CI: –32.91 – 6.19, p = 0.18). This is particularly notable because while eight participants in the same-page condition had to re-answer one or both items in the two-step measure to make their corrections, no participants in the different-page condition had to do so.

Taken together, the evidence suggests that presenting the two-step measure on separate pages yields fewer corrections, particularly in the general population. Speeding

[20] Here, our analytic sample comparison excludes one respondent from the different-page condition who spent 56 minutes on the survey. Including this individual increases the timing difference across conditions to five minutes. However, the timing difference does not meet the conventional threshold for statistical significance in any of our tests (i.e., whether we include this individual alone or along with the 115 NORC-excluded cases).

may explain this pattern fully – even if same-page formatting eases participant understanding of what is being asked, speeding effects may outweigh the positive effect of same-page formats. Alternatively, one intuition behind the need for same-page presentation may no longer be true. English-speaking cisgender adults may have become more familiar with the distinctions between sex assigned at birth and current identity, making separate questions less confusing than in earlier studies.

4.3 Response order

The final design choice we examine is the order of response options for both items of the two-step measure. We tested four formats: alphabetical, high-to-low population prevalence, low-to-high population prevalence, and random. Across all analyses, we find no evidence that ordering affects the distribution of responses, the prevalence of write-ins, or the number of corrections (Table 4).

In the LGBT sample, the frequency of selecting transgender as a response ranged from 4% in the alphabetical condition to 5% in the high-to-low prevalence condition, and the overall distribution did not differ significantly by response order (X^2 (9, n = 2,402) = 6.7, p = 0.67). Write-ins followed a similar pattern, with the fewest in the alphabetical condition (4%) and the most in the high-to-low condition (6%), though again the differences were not distinguishable from zero (X^2 (3, n = 2,402) = 2.1, p = 0.56). When looking at true corrections, the fewest (15%) occurred in the alphabetical condition and the most (30%) occurred in the high-to-low prevalence and random conditions (X^2 (3, n = 2,402) = 1.2, p = 0.76). Appendix Table A-9 shows the complete set of ordering results for the LGBT sample, including robustness checks by demographic subgroups, which do not reveal systematic differences by response order.[21]

To address the potential for primacy effects, specifically, we grouped conditions by whether the first response option was the same (see Appendix Table A-10). This also did not yield statistically significant differences on any of our outcomes. For example, transgender responses were not most frequent when that category appeared first in the current-identity item (e.g., in the low-to-high prevalence or some random conditions), nor were man/male responses most frequent when man/male appeared first (e.g., in some random conditions or in alphabetical with gender terms). Further, when looking at sex assigned at birth, the total number of true corrections to the sex-assigned-at-birth item did not differ depending on whether female or male appeared first (X^2 (1, n = 2,402) = 0.2, p = 0.655). For current identity, true corrections were somewhat more common when

[21] A few of the demographic subgroup robustness checks yield conventionally significant results using chi-square tests (e.g., for respondents in the Northeast and people whose religion was not listed), but those results do not hold when we apply Fisher corrections to account for small cell sizes.

woman/female appeared first (seven corrections) compared to when man/male or transgender appeared first (two corrections each respectively), but the difference across conditions was not statistically significant at the 0.05 level (X^2 (2, n = 2,402) = 1.5, p = 0.47).[22]

Table 4: Main effects of response ordering (Experiment 3), US LGBT sample

Outcome measure		*Alphabetical*	*Low to high*	*High to low*	*Random*	*p-value*
	Gender identity distribution (%)					
4-category identity distribution	woman/female	57%	55%	53%	52%	0.67
	man/male	35%	35%	37%	39%	
	transgender	4%	4%	5%	5%	
	different term/write-in	4%	5%	6%	5%	
2-category identity distribution	all other categories (woman, man, transgender)	96%	95%	94%	95%	0.56
	different term/write-in	4%	5%	6%	5%	
	Total (*n*)	590	610	619	583	
	Correction distribution (n, ref cat = no corrections)					
Corrections to either sex at birth or current identity	all corrections	12	9	11	12	0.86
	true corrections	3	5	6	6	0.76 (0.77)
	false corrections	10	4	6	6	0.36 (0.40)
	Total	590	610	619	583	
Corrections to either sex at birth or current identity for noncisgender participants only	all corrections	5	5	7	5	0.97 (0.97)
	true corrections	3	4	6	2	0.80 (0.81)
	false corrections	3	1	2	3	0.51 (0.50)
	Total	54	70	79	57	

Notes: The *p*-values for chi-square tests are shown for all comparisons. The *p*-values from Fisher's exact tests are shown in parentheses for all analyses where at least one expected cell count falls below five.

Analysis of the general population sample results in the same conclusion: None of the three outcomes vary by response order (see Appendix Table A-11).[23] Although null results in this case are not discouraging – they suggest comparability is not compromised by differences in response order for smaller surveys – finding no difference in the response distribution or these specific data quality metrics does not imply that order is inconsequential. Beyond statistical variation in the results, question design also carries symbolic weight, a point we return to below.

[22] To detect an effect of the observed magnitude ($\omega \approx 0.025$) would require a sample size of nearly 13,000, though an effect of that size is generally considered negligible.

[23] The fewest true corrections in the general US adult sample occurred in the high-to-low condition (0) while the most occurred in the low-to-high condition (3) and then the alphabetical condition (2), the opposite direction from the LGBT sample findings. A power calculation based on the observed effect size for true corrections ($\omega \approx 0.05$) indicated that a sample size of between 4,000 and 5,000 is needed to distinguish a difference of this magnitude, though an effect of that size is generally considered negligible.

5. Discussion

A large body of scholarship has made clear that survey researchers should use a two-step measure that asks separately about sex assigned at birth and gender identity to avoid conflating sex and gender or erasing transgender and nonbinary people in scientific research. To further advance inclusive data collection efforts, we examined three understudied design choices in implementing the two-step measure: whether to use sex or gender terms in the identity item, whether to present both items on the same page or on different pages, and how to order the response options. Based on our findings, we recommend that researchers working with samples of English-speaking adults should (1) use gender terms for the identity item, (2) present the two items on separate pages in online surveys, and (3) order sex and gender response options by population prevalence (see Figure 3 for an illustration of the proposed formatting). We address additional considerations related to these decisions below, including offering evidence that bears on our other design recommendations, such as whether to include a "nonbinary" identity response option and when (or of whom) confirmation questions should be asked.

Figure 3: Recommended two-step measure formatting

1. What sex were you assigned at birth, on your original birth certificate?

Female
Male
[Prefer not to say]

----------page break----------

2. What is your current gender?

Woman
Man
Transgender
Nonbinary
I use a different term (please specify):_____
[Prefer not to say]

Notes: Our study did not include a "prefer not to say" option. Rather, participants could choose to skip either of the two-step questions without providing an answer. In surveys where participants are forced to provide an answer to these questions, we recommend including the option "prefer not to say" (see NASEM 2022).

5.1 Identity response recommendations

The 2022 NASEM report recommended sex terms as response options for the identity item, in part because of a lack of contrary evidence, and because its recommendations were geared toward general population surveys. Nevertheless, our experiment shows that an identity item with gender terms performs better for an LGBT-focused sample of US adults and no worse among the general population. Using gender terms reduces write-ins, which can be hard to analyze and result in miscounting, as well as corrections, strengthening data quality. We therefore recommend pairing sex terms with the sex-assigned-at-birth item and gender terms with the current-identity item in surveys of English-speaking adults.[24] This recommendation is particularly strong for LGBT-focused studies and general population samples smaller than 12,000. Larger general population samples should consider testing whether the patterns we observe replicate before proceeding.

Although not experimentally tested, our data also highlight the importance of a dedicated nonbinary option for identity responses. In both samples, nearly half or more of all write-ins referenced nonbinary identities. For example, among the LBGT sample, writing in "nonbinary" or a related variation (such as "nonbinary trans," "genderfluid/nonbinary," or "nonbinary/transmasc") accounted for 45% of the open-ended responses.[25] Four-fifths (80%) of the write-ins in the general population study were "nonbinary" or a variation. Adding this category would reduce respondent burden and minimize the analytic challenges associated with coding open-ended responses. Of course, a free-text option still should be retained, as recommended by NASEM (2022), to allow for respondent autonomy and to monitor changing terminology. Our recommendation to add a nonbinary response is particularly relevant in surveys with younger respondents, among whom nonbinary identities are increasingly prevalent (Ridgeway and Saperstein 2024). The choice of whether to include a nonbinary response in addition to or instead of a transgender response (given that the two-step measure allows researchers to tabulate noncisgender identities) requires further study.

[24] Our sample, and thus conclusions, are limited to adult participants, but the two-step measure has been tested among youth participants in both clinical settings (Lau et al. 2021) and foster-care settings (Wilson et al. 2016) and was found to perform well in both. If current gender identity is asked in youth surveys, gender terms likely will need to be altered to "girl" and "boy" instead of, or alongside, "woman" and "man."

[25] The next most frequent terms in the LGBT sample included variations of gender fluidity or nonconformity, such as "genderfluid" (13.5%), "gender nonconforming" or "genderqueer" (6%), and "agender" (6%). The content of the write-ins did not differ by sex or gender terms condition, though a few write-ins did seem to match the terminology provided: For example, "transsexual" and "intersex" appeared in the sex terms condition while "trans man" and "trans woman" appeared in the gender terms condition. Only 3 of the 118 total write-ins in the LGBT sample could be viewed as "protest" write-ins. Two responses could be seen as protests against gender identity questions in general: "too old" and "am just a biological woman, I don't have interest in this." One response was a specific protest against using sex terms when asking about gender identity ("Woman is the proper gender term. Female is a sex.").

5.2 Pagination recommendations

Prior research supported competing expectations regarding the effect of page formatting on two-step measure responses: A study specific to the two-step measure suggests same-page formatting would improve comprehension for both cisgender and transgender respondents (Bauer et al. 2017), while generic survey research guidelines point to the benefit of putting each question on a separate page to reduce speeding through responses (Toepoel, Das, and Van Soest 2009). Our results did not reveal meaningful differences in response distribution for the LGBT sample. However, all corrections in the general adult sample occurred in the same-page condition. The balance of evidence points to the different-page format as better performing, in part because it did seem to encourage participants to slow down.

Because our study lacked item-level timing data, future research should explore this issue further. Ideally, this work should oversample groups prone to mis-clicks (reporting errors), capture item-level timing data, and conduct cognitive interviews with participants. Participants prone to mis-clicking may be difficult to identify a priori but could include people who tend to complete surveys more quickly than average, as well as people more likely to struggle with comprehension (e.g., older respondents or people with less education). Cognitive interviews could focus on whether comprehension of the two-step measure has improved among cisgender adults since the original studies were completed a decade or more ago and whether transgender respondents' reactions to the page format also may have changed. Such research could further consider how the order of the two-step measure (i.e., whether gender identity is asked before or after sex assigned at birth) interacts with the use of different-page placement in LGBT-specific and general population surveys. That said, our data suggest that defaulting to different-page formatting (given its common use in online surveys) should improve overall data quality for the two-step measure. This recommendation is strongest for general population surveys and smaller LGBT-focused studies.

5.3 Response order recommendations

We next turn to a discussion of the results from the response order experiment. Encouragingly, for researchers who compare multiple smaller surveys with different response option orders, we find no statistical differences in any outcome across our four ordering conditions. However, in general population samples with more than 4,000 participants, researchers should test whether significant differences in data quality emerge based on ordering decisions. Additional testing is warranted because ordering carries symbolic meaning.

Based on our findings, to avoid reinforcing patriarchal systems, and in line with the NASEM (2022) report, we recommend listing female and woman first, in the sex-assigned-at-birth and current-identity questions, respectively (see Figure 3); this ordering corresponds with high-to-low population prevalence in the United States and other English-speaking countries. Most, but not all, demographic questions currently order categories from high to low expected population size (e.g., race, religion) under the assumption that connecting well-known primacy effects with the most common response category will reduce measurement error.

It is worth noting that, descriptively, in our in our LGBT sample, the fewest write-ins and true corrections occurred in the alphabetical ordering (though statistically there was no detectable difference in write-ins and true corrections across the experimental conditions). However, we chose not to recommend alphabetical ordering for the two-step gender measure because doing so somewhat conflicts with our recommendation to use gender terms for the identity item, which is supported by the strongest evidence for differential item performance across our experimental conditions. In combination, those two formatting decisions would result in "female" appearing first for sex at birth and "man" appearing first for current identity (for an illustration of this issue, see Figure 1). This could increase mis-clicking, especially among respondents who speed through surveys and may, at least initially, expect to see the options in the same order (or have grown used to seeing "male" first, despite the lack of scientific justification for that ordering).

We also caution that consistency across items may matter as much as the specific order chosen, and the response order likely should be consistent across demographic items, especially when they are asked one right after another. This matters because the symbolic meaning of high-to-low expected population size varies: For the two-step measure putting "female" and "woman" first disrupts existing hierarchies; for sexual orientation or racial identity, where "straight/heterosexual" or "white" would appear first, it does not. Future research should go beyond quantitative metrics of data quality and consider how respondents interpret these response-ordering decisions and the consequences of (potentially) reinforcing inequality.

5.4 (Re)considering confirmation questions

Finally, though not experimentally manipulated in this study, we discuss the implications of our results for the inclusion of confirmation questions. Evidence from our study raises concern about how respondents interpret confirmation questions. Large-scale national surveys in the United States, such as the HPS and NCVS, have asked confirmation questions only of respondents whose initial responses suggest a noncisgender identity,

aiming to reduce false positives. This approach not only overlooks potential false negatives but also places disproportionate survey burden on gender minorities. In contrast, in this study, we asked all respondents a confirmation item. We find an unexpected number of "false" corrections, where respondents – perhaps unfamiliar with being asked to confirm their prior responses – requested a correction but ultimately affirmed their response. Among the true corrections, we see differing patterns between our two samples (Table 5). In the general population sample, where 8 total corrections were made, 75% were false positives. However, in the LGBT sample, where noncisgender identities were more common overall, the 44 total corrections included an equal number of false positives and false negatives (11% each).[26] Thus, after adjusting for corrections, the overall counts of cisgender and noncisgender responses in the LGBT sample remained the same.

Table 5: Correction type by sample

	Total number of corrections	*Non-cis → Cis ("false positive")*	*Cis → Non-cis ("false negative")*	*Cis → Cis*	*Non-cis → Non-cis*
US LGBT sample	44	11%	11%	39%	39%
US adult sample	8	75%	0%	13%	13%

Notes: The final two columns count instances where the cisgender classification remained stable (e.g., moving from noncisgender to noncisgender). Some of these corrections are false corrections, others are true corrections. For example, someone could remain noncisgender despite making a correction if they identified as transgender and only corrected their sex assigned at birth.

These results both reveal the added cognitive burden produced by asking respondents to confirm prior answers and cast doubt on the underlying justification for doing so. First, the direction and magnitude of potential errors (i.e., false positives and false negatives) may not align as expected, particularly in samples of gender and sexual minorities. Second, although false positives are a concern if they reflect measurement error (e.g., cisgender people who mis-clicked a different binary gender than their sex assigned at birth), it also is possible that asking for confirmation is instead taken as a sign of social desirability, as if asking someone, "Are you sure you want to identify as transgender or nonbinary?" Future research should interrogate this possibility directly, through cognitive interviews or more focused experiments.[27] In the meantime, we encourage researchers to carefully weigh the need for this type of questioning, especially

[26] The remaining corrections did not affect whether the respondents would be classified as cisgender (39%) or not (39%). For example, someone could remain noncisgender despite making a correction if they identified as transgender in the current-identity question and only corrected their sex assigned at birth.

[27] In our general adult sample, which included a final question asking for open-ended feedback about the survey, several cisgender respondents volunteered negative or surprised reactions to the confirmation questions. For example, a young woman said: "It was annoying to get the questions double checking what I said." A middle-aged man expressed similar concern: "I have never taken a study where the survey repeats my answer and asks, 'are you sure?'"

when only directed at already stigmatized minority identities, as it may cause more harm than good. If accuracy benchmarking is especially important for a given study, for example in a large-scale general adult survey used to produce official population estimates, researchers should consider randomly assigning a confirmation question to a subset of both cisgender and noncisgender responses to produce more comprehensive correction estimates, which – as our data suggest – may or may not represent mis-clicking "errors."

6. Conclusion

Taken together, our findings provide practical guidance for implementing the two-step gender measure. Using gender terms for the gender identity item, placing the two items on separate pages, and adopting a consistent ordering scheme will strengthen data quality without compromising comparability. Including a nonbinary response option and reconsidering when to deploy a confirmation question can further improve inclusivity in surveys of English-speaking adults.

Although some open questions remain or were newly raised by our experimental results, researchers should have confidence in using the two-step gender measure. The overall approach is well validated, and the formatting choices tested in this study represent minor refinements rather than fundamental changes. If anything, our findings suggest small differences in the formatting or design of the two-step measure do not compromise comparability or reliability, even as certain choices can improve inclusivity, sensitivity, and data quality. Researchers should weigh the considerations presented in this study carefully, especially in relation to their target sample size, but they can also be assured that the measure performs reliably across a range of subgroups and formats.

Our recommendations are especially timely in light of current controversies over how to best measure gender in surveys. The recent US presidential executive order has had wide-reaching effects on federally funded surveys and grant-supported projects. For example, the American Community Survey (ACS), administered by the US Census Bureau, reversed a Biden-era decision to test the inclusion of SOGI items, reverting instead to a binary sex question. Meanwhile, recent national estimates suggest that about 7% of US adults identify as LGBT – a proportion that has doubled in just the past decade and reaches as high as 20% among adults between 18 and 25 (Jones 2022). In response to this growing share of the population, SOGI questions, though rare, were increasing, in US flagship national surveys and other federal data collection (NASEM 2022). Following the executive order, gender identity questions are again absent altogether, leaving key dimensions of people's identities and experiences unmeasured (Bouton and Redfield 2026). By using our suggested formatting of the two-step gender measure, whenever

possible, researchers not only can help to fill long-standing gaps in knowledge about sex and gender, but also can be confident their demographic data are reflective of the diversity of the population.

7. Acknowledgments

The LGBT sample was supported through a special competition for targeted samples sponsored by Time-Sharing Experiments for the Social Sciences, NSF Grant 0818839, Jeremy Freese and James Druckman, Principal Investigators. We are also grateful to Tommy Ren and Elizabeth Deneen for research assistance; to the staff at NORC/AmeriSpeak, especially Suzanne Howard, for managing data collection; and to members of the Stanford Gender and Gender Inequality Workshop for their comments and suggestions.

Tessa Holtzman: Writing – original draft, Writing – review and editing, Methodology, Formal analysis, Data curation, Project administration.

Aliya Saperstein: Conceptualization, Writing – review and editing, Methodology, Formal analysis, Funding acquisition, Supervision, Project administration.

References

Ainsworth, C. (2015). Sex redefined. *Nature* 518: 288–291. doi:10.1038/518288a.

Alexander, A.C., Bolzendahl, C., and Wängnerud, L. (2021). Beyond the binary: New approaches to measuring gender in political science research. *European Journal of Politics and Gender* 4(1): 7–9. doi:10.1332/251510820X16067519822351.

Bauer, G.R., Braimoh, J., Scheim, A.I., and Dharma, C. (2017). Transgender-inclusive measures of sex/gender for population surveys: Mixed-methods evaluation and recommendations. *PLOS ONE* 12(5): e0178043. doi:10.1371/journal.pone.0178043.

Biggs, M. (2026). Comparing transgender identities in the census of Scotland and the census of England and Wales. *The British Journal of Sociology* 77(1): 163–169. doi:10.1111/1468-4446.70030.

Bornatici, C., Felder, M., Gianettoni, L., Mordasini, R., and Steinmetz, S. (2025). Measuring assigned sex, gender identity, and sexual orientation in population surveys. *FORS Guides* 26(1): 1–44. doi:10.24449/FG-2025-00026.

Bouton, L. and Redfield, E. (2026). Removal of sexual orientation and gender identity from federal data collections: January 2025 to January 2026. Los Angeles: The Williams Institute.

Cowan, S. (2005). 'Gender is no substitute for sex': A comparative human rights analysis of the legal regulation of sexual identity. *Feminist Legal Studies* 13(1): 67–96. doi:10.1007/s10691-005-1457-2.

Cronin, R.M., Jerome, R.N., Mapes, B., Andrade, R., Johnston, R., Ayala, J., Schlundt, D., Bonnet, K., Kripalani, S., Goggins, K., Wallston, K.A., Couper, M.P., Elliott, M.R., Harris, P., Begale, M., Munoz, F., Lopez-Class, M., Cella, D., Condon, D., AuYoung, M., Mazor, K.M., Mikita, S., Manganiello, M., Borselli, N., Fowler, S., Rutter, J.L., Denny, J.C., Karlson, E.W., Ahmedani, B.K., and O'Donnell, C.J., Vanderbilt University Medical Center Pilot Team, and the Participant Provided Information Committee (2019). Development of the initial surveys for the All of Us Research Program. *Epidemiology* 30(4): 597–608. doi:10.1097/EDE.0000000000001028.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80(1): 27–38. doi:10.1093/biomet/80.1.27.

Galesic, M., Tourangeau, R., Couper, M.P., and Conrad, F.G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly* 72(5): 892–913. doi:10.1093/poq/nfn059.

Guyan, K. (2022). *Queer data: Using gender, sex and sexuality data for action*. London: Bloomsbury Academic. doi:10.5040/9781350230767.

Hammarström, A. and Annandale, E. (2012). A conceptual muddle: An empirical analysis of the use of 'sex' and 'gender' in 'gender-specific medicine' journals. *PLOS ONE* 7(4): e34193. doi:10.1371/journal.pone.0034193.

Jones, J. (2022). LGBT identification in U.S. ticks up to 7.1% [electronic resource]. Gallup. https://news.gallup.com/poll/389792/lgbt-identification-ticks-up.aspx.

Korolczuk, E., Graff, A., and Kantola, J. (2025). Gender danger. Mapping a decade of research on anti-gender politics. *Journal of Gender Studies* 34(5): 621–640. doi:10.1080/09589236.2025.2489584.

Lau, J.S., Kline-Simon, A., Sterling, S., Hojilla, J.C., and Hartman, L. (2021). Screening for gender identity in adolescent well visits: Is it feasible and acceptable? *Journal of Adolescent Health* 68(6):1089–1095. doi:10.1016/j.jadohealth.2020.07.031.

Lindqvist, A., Sendén, M.G., and Renström, E.A. (2021). What is gender, anyway: A review of the options for operationalising gender. *Psychology and Sexuality* 12(4): 332–344. doi:10.1080/19419899.2020.1729844.

Loveman, M. (2014). *National colors: Racial classification and the state in Latin America*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199337354.001.0001.

Manfreda, K.L., Batagelj, Z., and Vehovar, V. (2006). Design of web survey questionnaires: Three basic experiments. *Journal of Computer-Mediated Communication* 7(3): JCMC731. doi:10.1111/j.1083-6101.2002.tb00149.x.

NASEM (2020). *Understanding the well-being of LGBTQI+ populations*. Washington, D.C.: National Academies Press. doi:10.17226/25877.

NASEM (2022). *Measuring sex, gender identity, and sexual orientation*. Washington, D.C.: National Academies Press. doi:10.17226/26424.

NORC (2023). TESS 123 Saperstein: Project methods and transparency report [electronic resource]. Chicago, IL: NORC at the University of Chicago. https://tessexperiments.org/study/saperstein1530.

Office of National Statistics (2021). Sex and gender identity question development for census 2021 [electronic resource]. UK Office of National Statistics. https://www.ons.gov.uk/census/censustransformationprogramme/questiondevelopment/sexandgenderidentityquestiondevelopmentforcensus2021.

Pao, C., Donnelly Moran, K., Compton, D.L., Kaufman, G., and Dowling, J.A. (2025). The case for 'other': Measuring gender and sexual identity in survey research. *Sociology Compass* 19(1): e70031. doi:10.1111/soc4.70031.

Restar, A.J., Lett, E., Menezes, N.P., Molino, A.R., Poteat, T.C., Dean, L.T., Glick, J.L., Baker, K.E., and Cole, S.W. (2024). Getting precise about gender and sex measurement: A primer for epidemiologists. *American Journal of Epidemiology* 193(12): 1861–1867. doi:10.1093/aje/kwae144.

Ridgeway, C.L. (2011). *Framed by gender: How gender inequality persists in the modern world*. New York: Oxford Press. doi:10.1093/acprof:oso/9780199755776.001.0001.

Ridgeway, C.L. and Saperstein A. (2024). Diversifying gender categories and the sex/gender system. *Annual Review of Sociology* 50: 385–405. doi:10.1146/annurev-soc-030222-035327.

Saperstein, A. (2022). *Stability of two-step sex and gender responses in U.S. panel data*. Paper presented at Population Association of America Annual Meeting, Atlanta, Georgia, April 6–9, 2022.

Saperstein, A. and Ren. H. (2026). *Optimal response ordering: Survey experimental evidence across identity items.* Paper presented at the American Association for Public Opinion Research, Los Angeles, CA, May 13, 2026. https://aapor.confex.com/aapor/2026/meetingapp.cgi/Paper/5277.

Saperstein, A. and Sobotka, T. (2025). *Gender conformity and wellbeing: An experiment in ordering effects*. Paper presented at the 30th International Population Conference, Brisbane, Australia, July 14, 2025. https://ipc2025.popconf.org/abstracts/250162.

Saperstein, A. and Westbrook, L. (2019). *Alternative gender measures survey*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. doi:10.3886/E109542V1.

Toepoel, V., Das, M., and Van Soest, A. (2009). Design of web questionnaires: The effects of the number of items per screen. *Field Methods* 21(2): 200–213. doi:10.1177/1525822X08330261.

Wells, T., Bailey, J.T., and Link, M.W. (2013). Comparison of smartphone and online computer survey administration. *Social Science Computer Review* 32(2): 238–255. doi:10.1177/0894439313505829.

West, C. and Zimmerman, D.H. (1987). Doing gender. *Gender and Society* 1(2): 125–151. doi:10.1177/0891243287001002002.

Westbrook, L. and Saperstein, A. (2015). New categories are not enough: Rethinking the measurement of sex and gender in social surveys. *Gender and Society* 29(4): 534–560. doi:10.1177/0891243215584758.

Wickes, R. and Emmison, M. (2007). They are all 'doing gender' but are they are all passing? A case study of the appropriation of a sociological concept. *The Sociological Review* 55(2): 311–330. doi:10.1111/j.1467-954X.2007.00707.x.

Wilson, B.D.M., Cooper, K., Kastanis, A., and Choi, S.K. (2016). Surveying LGBTQ youth in foster care: Lessons from Los Angeles. Los Angeles: The Williams Institute.