Comparing Select-One Versus Select-All-That-Apply Response Options in a Two-Step Gender Identity Question October 2023

Victoria R. Narine[1], Robin Kaplan[1], Rebecca L. Morrison[1]

[1]Bureau of Labor Statistics, 2 Massachusetts Avenue, NE, Washington, DC 20212

Abstract
Measuring sexual and gender minority (SGM) populations in Federal surveys improves understanding of these populations and can inform policy makers on disparities between those who identify as gender minorities and those who do not. Current understandings of SGM populations are dynamic and requires survey methodologists and statisticians to adapt to social shifts by maintaining an ongoing research agenda. To contribute to the expanding literature on the measurement of gender identity, these experiments explore (1) the potential effects of select-one versus select-all-that-apply response formats in a two-step gender identity series, where respondents are asked for their sex assigned at birth in one question, then their current gender identity in a separate question, and (2) the effects of the presence of a write-in box. Participants were recruited from a non-probability online panel and completed a self-response survey in Qualtrics (n=1362). In a 2 x 2 (Response Format x Presence of Write-In) factorial design, participants were randomly assigned to either the select-one or select-all-that-apply condition, and either the write-in present or no write-in present condition. Results suggest that neither response format nor the presence of a write-in box affect response distributions in a mostly cis-gender sample. However, future research should intentionally recruit gender minorities to better understand their question response behaviors. Implications, recommendations, and future research are discussed.

Key Words: question response format, select-all-that-apply, write-in, gender, experiment, response distribution, gender minorities, measurement

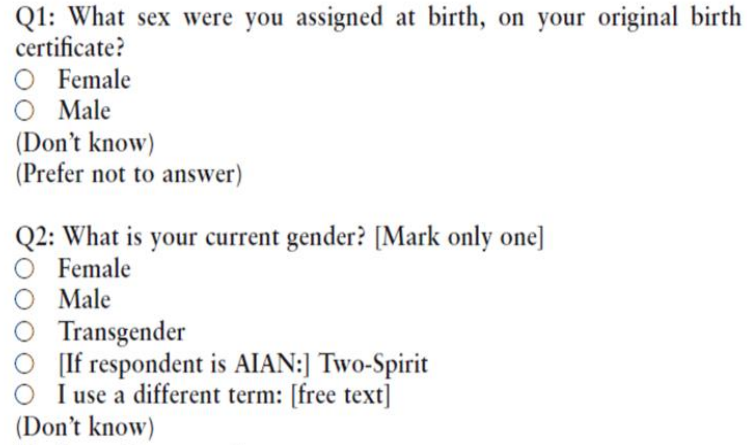
1. Introduction

The Federal Committee on Statistical Methodology's Measuring Sexual Orientation and Gender Identity (SOGI) Research Group developed an extensive research agenda to investigate current and best practices for measuring sexual and gender minorities (FCSM, 2016). The purpose of the research agenda is to promote continued improvement and adoption of SOGI measures in the federal government through inter-agency collaboration, shared lessons learned, and the research and development of SOGI survey items. The report put forth by the Research Group identifies three research priorities as important areas of focus when studying methodological issues around collection of SOGI data: (1) question terminology related to sex and gender identity, (2) proxy reporting, or whether respondents have the knowledge to report SOGI information about other members of their household, and (3) translation of SOGI concepts and/or questions into other languages.

Several additional factors guide the research and collection practices for sexual orientation and gender identity by federal agencies. First, Executive Order 14075 on *Advancing Equality for Lesbian, Gay, Bisexual, Transgender, Queer, and Intersex Individuals* (White House, 2022) broadly addressed policy and social issues affecting SGM populations. Specifically, the executive order calls for the advancement of inclusive and responsible data collection practices in SGM populations. Second, the National Academies of Sciences, Engineering, and Medicine issued a report that evaluated existing SOGI measures and provides recommendations for standardizing federal data collection efforts to minimize measurement error while promoting responsive and culturally sensitive SOGI questions (NASEM, 2022). Third, the Federal Evidence Agenda on LGBTQI+ Equity report provides guidance and direction to federal agencies developing their own SOGI data action plans (White House, 2023). Finally, an Office of Management and Budget (OMB) report documents the best practices and considerations for the collection of SOGI data when conducting federal statistical surveys (OMB, 2023). The current paper focuses on gender identity research.

To date, research across federal agencies has explored best practices for the collection of gender identity data using both quantitative and qualitative methods. This research has resulted in the consensus that a two-step approach to measuring gender identity is effective, used most often, and with the least amount of measurement error (see Figure 1). Importantly, the two-step approach distinguishes between sex assigned at birth and gender identity, with the sex assigned at birth question asked first (e.g., "What sex were you assigned at birth, on your original birth certificate")

followed by a broad measure of gender identity (e.g., "What is your current gender?")[1]. This approach is used by most agencies that collect gender identity data and was recommended by the National Academies report (NASEM, 2022), though the report also acknowledges the benefits and drawbacks of the questions within the two-step approach.


Q1: What sex were you assigned at birth, on your original birth certificate?
○ Female
○ Male
(Don't know)
(Prefer not to answer)

Q2: What is your current gender? [Mark only one]
○ Female
○ Male
○ Transgender
○ [If respondent is AIAN:] Two-Spirit
○ I use a different term: [free text]
(Don't know)

Figure 1: Recommended 2-Step Gender Identity Question (NASEM, 2023)

Although a two-step approach to measuring gender identity is the most frequently used approach by federal agencies, there still exists room for continued research and experimentation. For example, there is variation in the recommended question wording, response options, and response format of the two-step question. There is also a lack of agreement around the effect of the option to write-in an alternative gender identity when a respondent selects "I use a different term." Moreover, it is unclear how these factors are affected by mode of administration. The current study adds to this growing body of research by exploring differences in the response distributions between select-all-that-apply and select-one response formats, and whether the presence of a write-in box when "I use a different term" was selected affected data quality and/or meaningfully contributed to gender terminology. The purpose of this paper is to summarize the findings of this exploratory study and offer directions for future research.

2. Background

As the social and political visibility of gender minority populations increase, so too does the need for high quality gender identity data. However, there are methodological issues around the collection of gender identity, including the terminology used to describe various gender identities, the effect of proxy response on data quality, response format, and whether the presence of a write-in box when "I use a different term" is selected affects response distributions or leads to "protest" responses – potentially unrelated write-ins in protest of being asked about gender identity, usually on ideological grounds. This research focused on the latter two topics – response format and the effect of the presence of a write-in box. Both research topics are underexplored, resulting in a lack of evidence around how response format and response options may affect data quality and response distributions.

2.1 Response Format

Determining the response format depends on the measurement goal, the type of question being asked (e.g., evaluations or judgments about target object, event-based classification), what is being measured, response dimensions (e.g., occurrence, judgment about class membership) and the mode of administration (Schaeffer & Dykema, 2020). For example, the measurement of attitudes and opinions typically uses agree/disagree questions (i.e., response categories are an ordered selection with bipolar agree-disagree categories). In other cases, the measurement goal is achieved

[1] Although there is a consensus on the use of a two-step approach to measuring gender identity, there is still variation in wording for both steps. This wording is taken from the National Academies of Sciences, Engineering, and Medicine report.

using a simple yes/no response format, such as when measuring the presence of a quality or if an event occurred at some point in time. In many cases, item-specific response categories are required to correspond to the underlying content dimension of the question (Hohne & Lenzner, 2017). For example, a survey question measuring employment status may have response categories that include "employed," "self-employed," or "unemployed."

On federal surveys, the item-specific response format is commonly used because most federal surveys measure characteristics of the United States population. The use of an item-specific question format allows for simple tabulation of these characteristics and the subsequent publication of their respective tables for public consumption and external data users. Survey designers may choose to employ several different types of response formats, though, including select-all-that-apply, forced-choice, and select-one. Choosing which response format to use is commonly a product of the mode of administration. For instance, select-all-that-apply is most compatible with Web-based, self-administered surveys because it allows the respondent to see and read the response category labels at their own pace and select multiple categories (Smyth, Dillman, Christian, & Stern, 2006). Select-all-that-apply, however, is considered awkward when administered over the telephone, limiting its use in this mode. Conversely, a forced-choice format – whereby respondents answer either "yes" or "no" to each item in the list – is commonly used in telephone-administered surveys but advised against on self-administered surveys because of acquiescence bias, or the tendency to endorse (select "yes") more items than are true (Callegaro, 2015).

An assumption undergirding the switch between select-all-that-apply for self-administered modes and forced-choice for interview surveys is that the two formats are interchangeable and produce similar survey estimates (Callegaro, et al., 2014; 2015; Smyth et al., 2006). Survey research literature suggests that this is not necessarily the case. One possible explanation for differences in estimates between the two modes is that forced-choice may require deeper cognition than select-all-that-apply (Neuert, 2020; Smyth et al., 2006). Whereas respondents satisfy the requirements of a select-all-that-apply format by selecting the first category that they feel is a minimally justifiable response, a forced-choice format requires elaboration and deeper processing of each item (Smyth et al., 2006). Neuert (2019) used eye-tracking to record participants' eye movements when answering a series of eight items formatted as either select-all-that-apply or forced-choice. Participants endorsed significantly more items in the forced-choice format than the select-all-that-apply format. Across all four experiments, participants took more time to complete the items when they were in forced-choice compared to select-all-that-apply formats, and they fixated longer on the forced-choice than select-all-that-apply format. These results lend evidence to the theory that more time and cognitive effort are invested in the completion of forced-choice questions than select-all-that-apply question formats.

Because the different response formats are associated with different cognitive strategies, it may be inappropriate to assume function equivalency across the two formats. In fact, several studies find that response behavior does differ between select-all-that-apply and forced-choice. In one study in which university students were randomly assigned to either a select-all-that-apply or a forced-choice format in two otherwise identical surveys on campus life, Smyth et al. (2006) found that forced-choice elicited significantly more endorsement than did select-all-that-apply. In a similar follow-up study, Smyth et al. (2008) replicated their previous results using web and telephone. In self- and interviewer-administered modes, respondents endorsed significantly more items in a forced-choice format than a select-all-that-apply format. In yet another study, Thomas and Klein (2006) demonstrated across five experiments that, on average, respondents selected significantly fewer select-all-that-apply categories than they endorsed for forced-choice items. Finally, in a series of meta-analyses, Callegaro et al. (2015) found that estimates from select-all-that-apply formats were consistently and reliably lower than estimates from forced-choice formats.

2.1.1 Response Formats of Gender Identity Measures

Given that select-all-that-apply has been shown to produce consistently lower estimates across randomized studies and may result in respondents failing to read the entire list of items and selecting every appropriate item, select-all-that-apply formats should typically be avoided in favor of other formats such as forced-choice (Dillman et al., 2014). However, selecting a response format may be complicated by the context of the question; that is, what is being measured. This may be especially true for highly salient topics and personal questions, like those that measure social identities. Despite recommendations to avoid select-all-that-apply, some survey research literature suggests that in situations where the topic is salient (e.g., social identity), using select-all-that-apply may be acceptable. This research suggests that topic salience may increase respondents' motivation to participate, leading to survey completion (Groves et al., 2004, 2000; Marcus et al., 2007). Additionally, increased topic salience may result in respondents reading each item carefully and completely, taking time to consider each response option, and selecting the best option (Holland & Christian, 2009).

Since social identities such as race, ethnicity, and gender are considered key components of one's self-concept, these social identities, then, are arguably salient for the respondent (Tajfel & Turner, 1979, 1986). Currently, many social identities such as race and ethnicity are measured using a select-all-that-apply format, allowing the respondent increased flexibility to define their social identity using as many nonexclusive categories as they see fit (Brenner & Bulgar-Medina, 2018). Little data supporting the use of select-all-that-apply formats exists, though. In one of the few studies, Brenner and Bulgar Medina (2018) tested whether topic salience influenced data quality in a predominantly lesbian, gay, bisexual, and queer (LGBQ) sample. Brenner and Bulgar-Medina (2018) conclude that asking about gender identity using a select-all-that-apply format works equally as well as in a forced-choice format among straight participants, suggesting that the format would work just as well in the general population. Brenner and Bulgar-Medina (2018) intentionally recruited sexual minorities, *not* gender minorities. This is a key caveat to the statement that select-all-that-apply is functionally equivalent to forced-choice.

2.2 Write-In Text Boxes in Gender Identity Measures

In general, survey designers and researchers avoid including an "other" category because of the risk of underreporting. Instead, response options should be as exhaustive as possible and include all reasonable alternative answers (Sudman and Bradburn, 1982). In the case of gender identity, however, including all possible categories as a response option is difficult, if not impossible, because of the multitude of terms used to describe one's gender identity. To avoid the potential for misclassification or misidentification, gender identity measures include an "other" or "something else" response category that is typically followed by a text-box that allows the respondent to freely identify themselves (FCSM, 2016). Research shows that the decision to include "other" or "something else" as a response option for sexual and gender identity questions may lead to the misclassification of these key social identities and negatively impact survey estimates (West & McCabe, 2021). Ridolfo, Miller, and Maitland (2012) found that the response option "other" acted as a "nonresponse category," whereby individuals who did not understand the response options or were unsure of how to respond accurately showed a propensity to select this category. Moreover, the respondents who selected "other" tended to be a more diverse subgroup defined by different sexual and gender identities. While the inclusion of the "other" response category and the write-in text box provided unique responses, it makes meaningful comparative analyses difficult. Eliason and Streed (2017) found similar results in their study, which included a "something else" category and a write-in box; in that study, the responses to "something else" were varied and made tabulations difficult. Other research has found that the responses to "something else" tend to come from individuals who are female, White, and younger (Feuer, 2022).

Omitting a "something else" category and the potential to write-in additional details poses its own risk. The increased heterogeneity of the subgroups when there are *fewer* response options (i.e., no "something else" response category and no option to write-in) may adversely impact population estimates. Specifically, predictor variables that are prone to misclassification errors (e.g., gender identity) and that are more heterogeneous than they would be with more response categories may result in biased estimates of the relationship between the predictor variable (e.g., gender identity) and other variables of interest (Savoca, 2000). In a study that analyzed public-use data from the National Survey of Family Growth (NSFG), West and McCabe (2021) assessed whether including or omitting a "something else" response category with the option to write-in an alternative sexual identity impacted population estimates. They found evidence that forcing respondents to select one of three sexual identities (i.e., heterosexual, gay, or bisexual) significantly and adversely affected population estimates compared to when a "something else" response option was offered. The authors concluded that – while possibly complicating tabulation efforts – the inclusion of "something else" more accurately depicts relationships between identity groups and the dependent variable of interest.

Another issue to consider is the potential for "protest" responses, or responses from individuals who are ideologically opposed to being asked about their gender identity or being presented with the possibility that gender may not be binary. In fact, some research has found that a significant portion of received write-in responses are not sexual and gender minorities identifying their preferred identities, but rather sexual and gender *majority* individuals "protesting" the question (Ridolfo et al., 2012; Truman et al., 2019). Research on write-in responses, though, is mixed. Researchers at the National Center for Education Statistics found that the proportion of ideological protests varied by study, ranging from 3.0% - 43.0% of the .3-1.0% of write-in responses (Richards et al., 2022). Other researchers have found that a significant majority of write-in responses comprise ideological "protest" responses, with many of these written responses conflating sexual orientation with gender identity or sexual majority individuals negatively reacting to the sexual orientation question (e.g., "Christian male"; Bates et al., 2019). Feuer et al. (2022) found that individuals who write "protest" responses tended to be older and male, and, in some cases, White and more educated.

Despite the possibility of ideological "protest" responses, NASEM (2022) recommends the inclusion of a write-in category as a strategy for providing individuals with a method of reporting the gender identity that best describes them. The two-step gender identity question recommended by NASEM (2022) uses "I use a different term" versus "something else" or "other" to refrain from marginalizing individuals whose identity is not captured by the available response options.

3. Research Questions

The current research adds to this growing body of research by comparing the effect of select-all-that-apply and select-one response options in gender identity measures on response distributions. Additionally, this research explored the effect of including a write-in text box versus not including a text box on the selection of the "I use a different term" option. This research was exploratory but was guided by four research questions. The remainder of this paper will summarize the methods and findings from these four research questions:

1. Do the response distributions of gender identity differ between select-all-that-apply and select-one?
2. Does a select-all-that-apply response format result in a higher number of categories selected compared to select-one?
3. Is there a difference in the time spent on the page for select-all-that-apply versus select one?
4. Does including a write-in text box when "I use a different term" is selected affect response distributions and/or lead to more "protest" responses?

4. Methods

3.1 Design and Materials

The goal of this study was to assess the impacts of select-all-that-apply versus select-one response formats and write-in text box versus no text box on a measure of gender identity. These independent factors were embedded in a survey and participants were randomly assigned to see one version of each factor.

The questions were administered online May 8 – May 15, 2023, as part of a larger screener questionnaire for research on two Bureau of Labor Statistics surveys as well as exploratory research on race and ethnicity. The questionnaire was administered on Qualtrics, an online survey administration platform. The instructions stated that the survey would take about five minutes and that they might be invited to participate in a follow-up survey based on their responses. Participants were first asked for their current employment status, whether they were employed by government, by a private company, a non-profit organization, or self-employed, and if they were a salaried worker, contract worker, or got paid by the hour. Participants who reported being unemployed received questions unrelated to gender identity before answering the two-step gender identity question. All other participants were directed to the gender identity questions after answering the initial employment questions. Because some participants were deemed ineligible for certain parts of the screener based on their responses to the employment question, the total number of questions received differed by participant. However, all participants received the 2-step gender identity question series immediately before the race and ethnicity questions, which were the last questions administered on the screener questionnaire.

All participants received the same sex assigned at birth question, "What sex were you assigned at birth, on your original birth certificate?" (See Figure 2.)

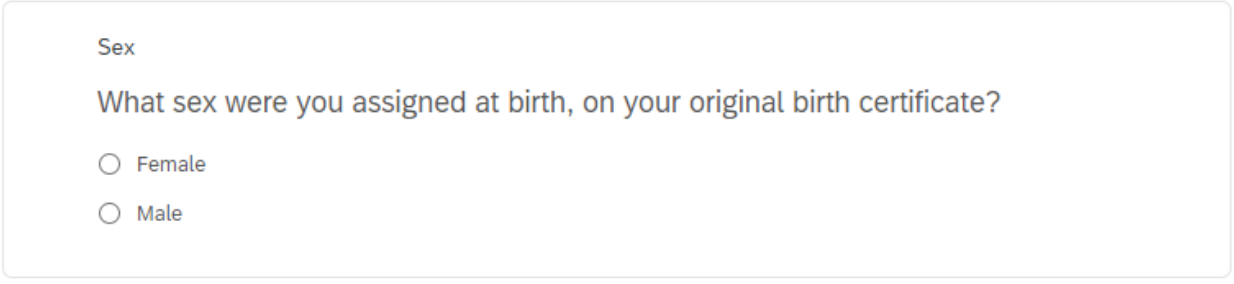


Figure 2. Step 1 of 2-step gender identity measure as seen by participants.

When participants received the gender identity question, they were randomly assigned to one of two response format conditions – select-all-that-apply versus select-one – and one of two write-in text box conditions – write-in text box versus no write-in text box (Figure 3). Participants assigned to the write-in text box condition were able to type in their response if "I use a different term" was selected.

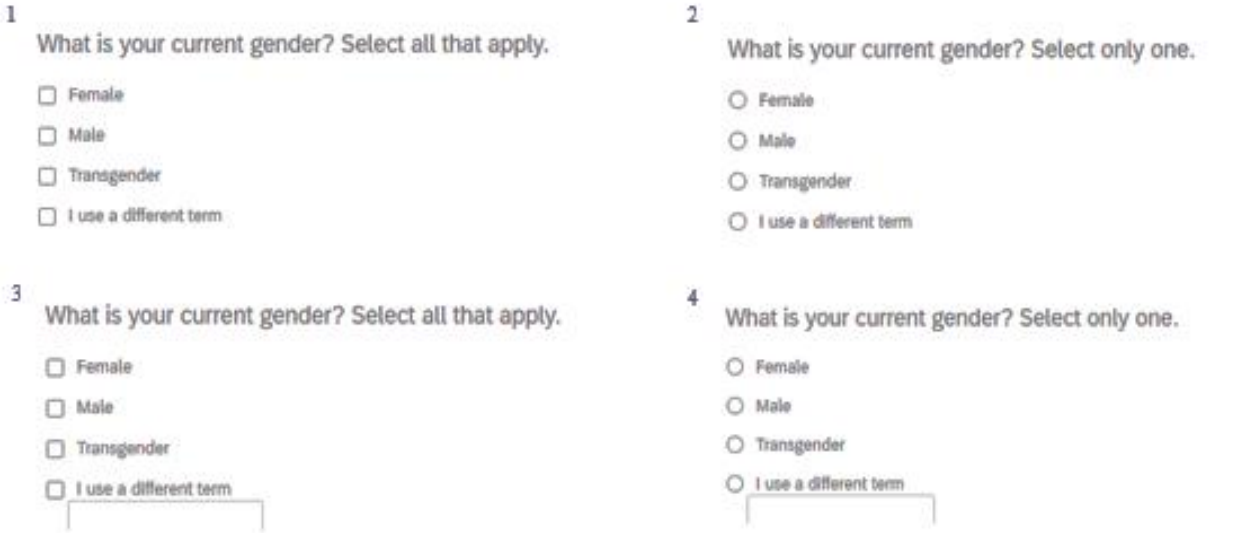


Figure 3. Step 2 of 2-step gender identity measure as seen by participants, by conditions. (1) Select-all-that-apply, no write-in; (2) Select-one, no write-in; (3) Select-all-that-apply, write-in; and (4) Select-one, write-in.

3.2 Participants

Participants were recruited from Amazon Mechanical Turk (mTurk), an online recruitment platform where individuals can complete short tasks (e.g., research studies) for pay. MTurk provides researchers with access to a large number of potential research participants. Though diverse, the mTurk population is not representative of the United States population (Paolacci & Chandler, 2014). Even so, MTurk is considered to be an efficient method of recruiting many individuals over a short period of time for research (Yu et al., 2019). Table 1 provides the demographic characteristics of the sample for this study.

1393 participants completed the questionnaire and the two-step gender identity question. As is expected in mTurk-recruited panels, the sample skewed younger, with 54.08% 40 years or younger, white (78.81%), and more educated, with the majority (68.75%) having an undergraduate degree or higher. Participants were also overwhelmingly employed, with 87.82% saying they were employed full-time[2].

[2] The employment question on the survey read "What is your employment status? Do not include work done on Amazon Mechanical Turk."

Table 1. *Demographic Characteristics of Participants (n=1393)*

Demographic	Category	Count (Percent)
Age	18-29	198 (14.53%)
	30-39	539 (39.55%)
	40-49	342 (25.09%)
	50-59	189 (13.87%)
	60+	94 (6.90%)
Race/Ethnicity[3]	White	1071 (78.81%)
	Hispanic or Latino	92 (6.77%)
	Black or African American	134 (9.86%)
	Asian	122 (8.98%)
	American Indian or Alaska Native	20 (1.47%)
	Middle Eastern or North African	0 (0%)
	Native Hawaiian or Other Pacific Islander	5 (0.59%)
Education	Less than high school	2 (.15%)
	High school graduate, no college	99 (7.26%)
	Some college or associates' degree	325 (23.84%)
	Bachelor's degree or higher	937 (68.75%)
Employment Status	Employed full-time	1197 (87.82%)
	Employed part-time	96 (7.04%)
	Self-employed or business owner	43 (3.15%)
	Not employed	27 (1.98%)

5. Results

4.1 Comparison of Gender Identity Response Distributions

Table 2 shows the frequency of each response combination – sex assigned at birth by gender identity. Only 0.01% of the sample identified as something other than cisgender. Cells with a dash (-) had a frequency of 0 such that no participant identified with that combination of sex assigned at birth and gender identity. The gender identity response distributions did not differ between the response format conditions, suggesting that response format may not impact estimates with a mostly cis gender sample.

Table 2. *Frequency of Each Response Combination by Condition*

	Select-One (*n* = 684)				Select-All-That-Apply (*n* = 679)			
Sex assigned at birth	Female	Male	Transgender	I use a different term	Female	Male	Transgender	I use a different term
Female	313 (45.76%)	1 (0.15%)	2 (0.29%)	2 (0.29%)	305 (44.92%)	2 (0.29%)	-	2 (0.29%)
Male	-	361 (52.78%)	3 (0.44%)	2 (0.29%)	2 (0.29%)	365 (53.76%)	2 (0.29%)	4 (0.59%)
Total	313 (45.76%)	362 (52.92%)	5 (0.73%)	4 (0.58%)	307 (45.21%)	367 (54.05%)	2 (0.29%)	6 (0.88%)

4.2 Comparison of Number of Categories Selected

One benefit of using a select-all-that-apply response format is the inherent flexibility that respondents have to select all response categories that best reflect the construct being measured. Since gender identity is socially constructed and individually defined, we explored whether participants utilized the ability to select multiple response options in the select-all-that-apply format. Across both conditions, the median number of categories selected was 1.0, suggesting

[3] The race/ethnicity question was a select-all-that-apply format.

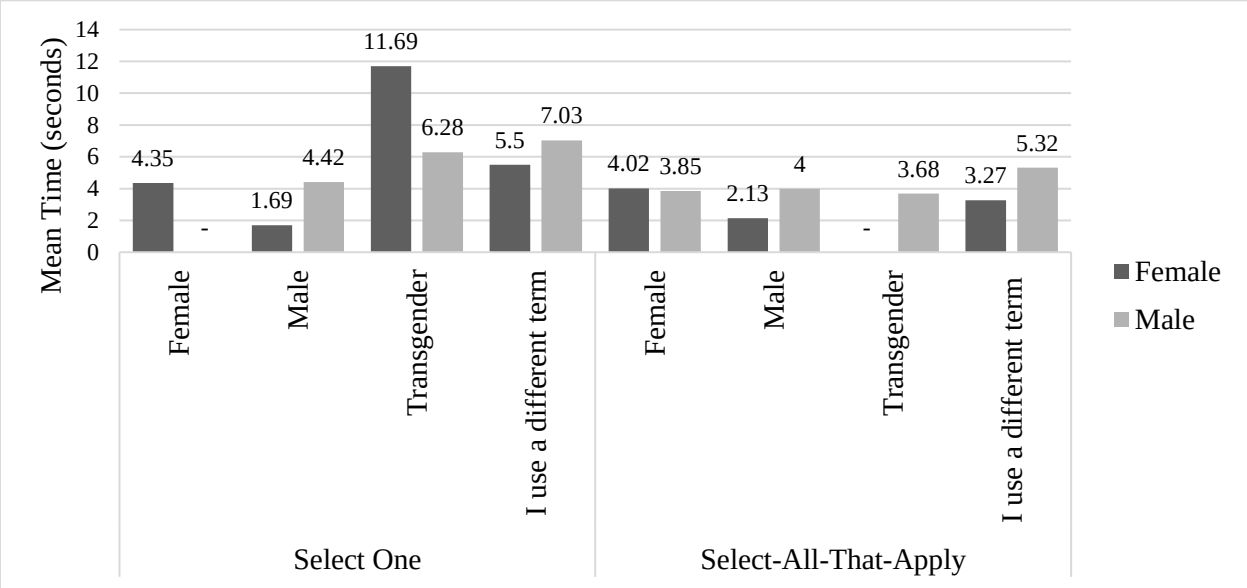
that participants did not need the flexibility of a select-all-that-apply format to accurately represent their gender identity.

To assess whether select-all-that-apply and select-one response options required different cognitive strategies, we analyzed the number of mouse clicks by condition and gender identity. We theorized that the greater the number of mouse clicks on the page, the greater the indecision. It may be that the rigidity of a select-one option for gender identity may result in more mouse clicks because gender minority participants have a more difficult time selecting a single option. Multiple mouse clicks are only a sign of indecision in the select-one condition since multiple mouse clicks would be expected in the select-all-that-apply condition. No differences between the two response format conditions were found (Table 3).

4.3 Comparison of Mean Time Spent on Page

When a survey question is difficult to answer, it is more cognitively burdensome and thus may take more time to answer (Jenkins & Dillman, 1997; Tourangeau, Rips, & Rasinski, 2000). We analyzed the time spent answering the gender identity question (operationalized as the time spent on the page before an answer was submitted) to determine if one response format was more cognitively burdensome than the other (Figure 4). An ANOVA found that how participants reported sex assigned at birth ($F_{3, 1359} = .37$, $p = .77$ or gender identity ($F_{3, 1359} = .12$, $p = .95$) had no significant relationship with time spent on page. Since there was no main effect, post hoc comparisons were nonsignificant. These results suggest that there is not a relationship between sex assigned at birth, gender identity, or response format on cognitive burden. However, the imbalance of responses among these categories may be obscuring the identification of differences in cognitive burden.

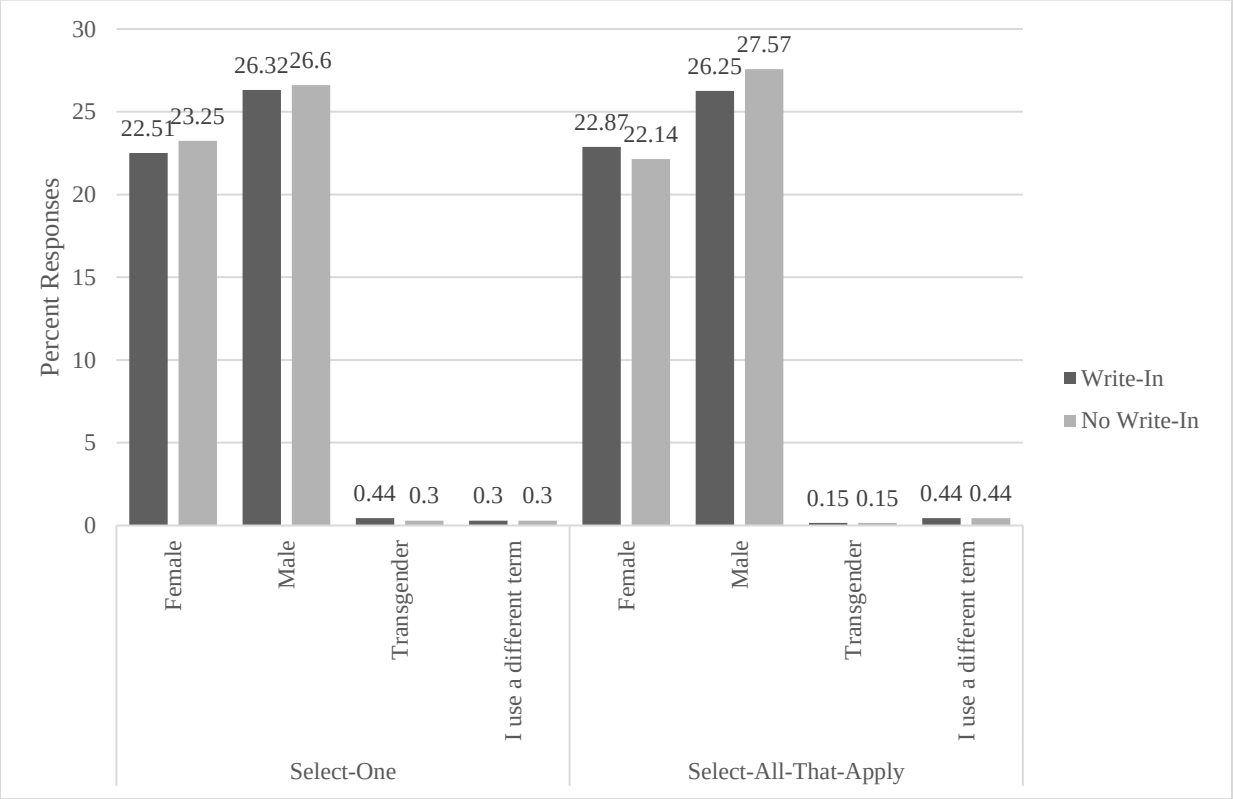
Figure 4. *Mean Time Spent on Gender Identity (Step 2) Page by Response Format and Gender Identity*



4.4 Comparison of Gender Identity Response Distributions by Write-In Condition

We compared gender identity response distributions across response format condition and write-in condition to assess whether the presence of the write-in text box with "I use a different term" affected how participants identify. It may be the case that respondents feel as if the available response categories are not exhaustive of their identity and would rather write-in their own category. We were also curious if the presence of a write-in text box would be more or less utilized by participants depending on the response format condition. Significance testing showed no effect of write-in condition on the distribution of gender identity responses, suggesting that seeing the write-in option on the screen does may not change estimates of gender identity in a mostly cisgender sample (Figure 5).

Figure 5. *Percent of Gender Identity Responses (Step 2) by Response Format and Write-In Condition*



4.4.1 Write-in Responses

Finally, we aggregated the responses provided in the write-in text box (when available) to assess whether the responses were meaningful or ideological "protest" responses. When the write-in text box was present, only 3 participants offered a response. Two of these 3 responses could be construed as meaningful additions to gender identity terminology: "Genderqueer" and "Non-binary." One response was a refusal to disclose. A fourth participant selected "I use a different term" in the write-in condition but left the text box empty.

6. Summary and Discussion

This research offers preliminary evidence that question response format does not affect the response behaviors of respondents in a mostly cisgender sample. First, this study investigated whether question response format impacted the response distributions of a two-step gender identity questions. Second, we explored whether response format affected the degree of cognitive burden, which was operationalized as the time spent answering the gender identity question (step 2 of the two-step question) before an answer was submitted and the number of clicks on a page before an answer was submitted. Third, we were interested in whether participants would select more responses options than in the select-one response format condition. And finally, we explored the effect of including a write-in text box when "I use a different term" was selected on response distributions and if the inclusion of a write-in text box resulted in ideological "protest" responses, or responses (typically by sexual and gender majority members) based on moral or ideological disagreement with being asked about gender identity or the possibility of gender being non-binary.

Results from this study indicate there may not be a difference in response distributions for a two-step gender identity question between select-all-that-apply or select-one response formats. There was no evidence of a relationship between response format and reported gender minority status. This study – with a predominantly cisgender sample -- also found no apparent difference in cognitive burden between the response format conditions. On average, participants spent as much time answering the gender identity question when the response format was select-one as they did when the response format was select-all-that-apply. Similarly, the number of mouse clicks on the page before an answer to the

gender identity question was submitted did not differ between the conditions. Moreover, participants did not select more response categories when the response format was select-all-that-apply compared to select-one. These findings suggest that participants may not need the flexibility provided by a select-all-that-apply format to accurately represent their gender identity. They also suggest that a select-one response format is not more cognitively taxing because of its rigidity compared to a select-all-that-apply format. Results from this study also indicate that the inclusion of a write-in text box when "I use a different term" is selected may not affect the response distribution, regardless of response format condition. Finally, the availability of a write-in text box did not result in ideological "protest" responses. This finding is particularly interesting since prior research has found that cisgender respondents typically comprise the majority of "protest" responses (Freuer et al., 2022).

Given that the sample was majority cisgender, conclusions regarding the response behaviors of gender minorities based on the current data are limited. Findings from this study offer preliminary evidence that response format for a two-step gender identity question does not affect response distributions, but this conclusion cannot be confidently stated without intentionally recruiting gender minorities. Thus, these results should be regarded as a first step in this line of research on the response behaviors of gender minorities. However, response format should still be considered when administering a two-step gender identity question. Providing respondents with the flexibility to select as many options as they feel best represent their identity could be self-affirming for those who identify as gender minorities and not impact data quality.

In addition to the small sample of gender minority individuals in the sample, there are other limitations that prevent us from confidently concluding that response format does not affect response distributions for a two-step gender identity question. This study drew from a non-probability sample that is not representative of the general population. The sample skewed White, young, and educated. This is typical of web panels but is not typical of the general population. Thus, the results are not necessarily generalizable.

To increase confidence in these results, future research should intentionally recruit gender minority individuals to better understand their response behavior. Another viable line of research is exploring how the addition of gender categories – like genderqueer, non-binary, or gender fluid – might affect what options someone chooses. Relatedly, research on response options should be ongoing. The language used to describe one's gender is not necessarily static and may change with the social landscape so the list of available response options may need updating as time passes. Relatedly, more research on measuring gender identity over time (i.e., in longitudinal surveys) should be conducted. The terms used to describe oneself may change as a function of age.

References

Bates, N., García Trejo, Y. A., & Vines, M. (2019). Are sexual minorities hard-to-survey? Insights from the 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS) Survey. *Journal of Official Statistics, 35*(4), 709-729.

Brenner, P. S., & Bulgar-Medina, J. (2018). Testing mark-all-that-apply measures of sexual orientation and gender identity. *Field Methods, 30*(4), 357-370.

Callegaro, M., Murakami, M. H., Tepman, Z., & Henderson, V. (2015). Yes-no answers versus check-all in self-administered modes: A systematic review and analyses. *Internal Journal of Market Research, 57*(2), 203-224.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail and mixed-mode surveys: The tailored design method* (4th ed.). Hoboken, NJ: John Wiley.

Eliason, M.J. and Streed, C.G. Jr. (2017). Choosing "something else" as a sexual identity: Evaluating response option on the National health Interview Survey. *LGBT Health, 4*, 376-379.

Federal Interagency Working Group on Improving Measurement of Sexual Orientation and Gender Identity in Federal Surveys. (2016). *Current measures of sexual orientation and gender identity in federal surveys*. Federal Committee on Statistical Methodology. https://nces.ed.gov/FCSM/pdf/buda5.pdf

Federal Interagency Working Group on Improving Measurement of Sexual Orientation and Gender Identity in Federal Surveys. (2016). *Toward a research agenda for measuring sexual orientation and gender identity in federal surveys: Findings, recommendations, and next steps*. Federal Committee on Statistical Methodology. https://www.fcsm.gov/assets/files/docs/SOGI_Research_Agenda_Final_Report_20161020.pdf.

Feuer, S., Virgile, M., Holzberg, J., & Ellis, R. (2022). *Who says that?: Demographic differences in open-ended responses to survey questions about sexual orientation.* [Paper presentation]. American Association for Public Opinion Research Conference, Chicago, IL.

Groves, R., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly, 68*(1), 2-31.

Groves, R., Singer, E., & Corning, A. (2000). Leverage-salience theory of survey participation: Description and an illustration. *Public Opinion Quarterly, 64*, 299-308.

Holland, J. L., & Christian, L. M. (2009). The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. *Social Science Computer Review, 27*(2), 196-212.

Karem Höhne, J., & Lenzner, T. (2017). New insights on the cognitive processing of agree/disagree and item-specific questions. *Journal of Survey Statistics and Methodology, 6*(3), 401-417.

Marcus, B., Bosnjak, M., Lindner, S., Pilischenko, S., & Schuetz, A. (2007). Compensating for low topic interest and long surveys: A field experiment on nonresponse in web surveys. *Social Science Computer Review, 25*, 372-383.

National Academies of Sciences, Engineering, and Medicine. (2022). *Measuring Sex, Gender Identity, and Sexual Orientation*. https://nap.nationalacademies.org/26424.

Neuert, C. E. (2019). Do forced-choice (FC) questions trigger deeper cognition than check-all-that-apply (CATA) questions? *Journal of Survey Statistics and Methodology, 8*(4), 617-635.

Office of the Chief Statistician of the United States. (2023). *Recommendations on the best practices for the collection of sexual orientation and gender identity data on federal statistical surveys*. Office of Management and Budget. https://www.whitehouse.gov/wp-content/uploads/2023/01/SOGI-Best-Practices.pdf.

Office of Management and Budget. (1995). *Standards for the Classification of Federal Data on Race and Ethnicity.* https://obamawhitehouse.archives.gov/omb/fedreg_race-ethnicity#:~:text=In%20decennial%20censuses%2C%20for%20example,data%20on%20ethnicity%20were%20collected.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*(3), 184-188. https://static1.squarespace.com/static/5197be40e4b0724ab77021be/t/544da65ee4b0dfe393bf7d6b/1414375006001/Paolacci+and+Chandler+2014.pdf.

Richards, D., Christopher, E., & Spiegelman, M. (2022, October 25 – 27). *Assessing open-ended self-*

reports of sexual orientation and gender identity: Is there room for improvement? [Paper presentation]. Federal Committee on Survey Methodology Research and Policy Conference, Washington, DC. https://www.fcsm.gov/assets/files/docs/2022-conference-docs/C5.4_Richards.pdf.

Ridolfo, H., Miller, K., & Maitland, A. (2012). Measuring sexual identity using survey questionnaires: How valid are our measures? *Sexuality Research and Social Policy, 9*, 113-124.

Schaeffer, N. C., & Dykema, J. (2020). Advances in the science of asking questions. *Annual Review of Sociology, 46*, 37-60.

Tajfel, J., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33-47). Monterey, CA: Brooks/Cole.

Tajfel, J., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (2nd ed., pp. 7-24). Chicago: Nelson-Hall.

Thomas, R.K., & Klein, J.D. (2006). Merely incidental? Effect on response format on self-reported behavior. *Journal of Official Statistics*, 22(2), 221-244.

Truman, J.L., Morgan, R. E., Gilbert, T., & Vaghela, P; (2019). Measuring sexual orientation and gender identity in the National Crime Victimization Survey. *Journal of Official Statistics, 35*, 835

Savoca, E. (2000). Measurement errors in binary regressors: An application to measuring the effects of specific psychiatric diseases on earnings. *Health Services and Outcomes Research Methodology, I* (149-164.

Smyth, J.D., Christian, L.M., & Dillman, D.A. (2008). Does "yes or no" on the telephone mean the same as "check-all-that-apply" on the web? Public Opinion Quarterly, 72(1), 103-113.

Subcommittee on Sexual Orientation, Gender Identity, and Variations in Sex Characteristics (SOGI) Data, & Subcommittee on Equitable Data. (2023). *Federal evidence agenda on LGBTQI+ equity*. National Science and Technology Council. https://www.whitehouse.gov/wp-content/uploads/2023/01/Federal-Evidence-Agenda-on-LGBTQI-Equity.pdf.

Sudman, S. & Bradburn, N. M. (1982). *Asking Questions*. San Francisco, CA: Jossey-Bass.

West, B. T., & McCabe, S. E. (2021). Choices matter: How response options for survey question about sexual identity affect population estimates of its association with alcohol, tobacco, and other drug use. *Field Methods, 33*(4), 335-354.

The White House. (2022, June 15). *FACT SHEET: President Biden to sign historic executive order advancing LGBTQI+ equality during Pride Month* [Press release]. https://www.whitehouse.gov/briefing-room/statements-releases/2022/06/15/fact-sheet-president-biden-to-sign-historic-executive-order-advancing-lgbtqi-equality-during-pride-month/.

Yu, E., Clark Fobia, A., Graber, J., Holzberg, J., Kaplan, R. L., Kopp, B., Kubzdela, K., Mockovak, W., Morrison, R.L., & Scanlon, R. (2019). *White paper: Experiences using online testing to support survey-methods research and pre-testing in the federal government*. https://www.bls.gov/osmr/research-papers/2019/pdf/st190010.pdf.