

# Self-identification of gender and sexual orientation in the US General Social Survey:

A review of methodological inconsistencies and anomalous data

Oliver Daniel & Tiago Martins

due March 20, 2022\*

## Abstract

The US General Social Survey, administered by the National Opinion Research Center at the University of Chicago, was recently emended to include options for respondents to specify several aspects of their sexual orientation and gender identity, including how their gender has changed relative to their birth sex. However, upon exploration of the data, it became clear that several major inconsistencies and anomalies existed in both the methodology by which the data were collected, as well as in the data themselves. This paper highlights these issues, and offers as an appendix a supplementary survey that would help to rectify these anomalies in future General Social Surveys.

## Contents

Introduction . . . . .	1
Data . . . . .	1
Results . . . . .	4
Discussion . . . . .	7
Survey . . . . .	7
References . . . . .	8

## Introduction

TODO

## Data

### Dataset

This paper focuses on the United States General Social Survey (US-GSS, or GSS in the the following), a survey regularly administered by the National Opinion Research Center (NORC) at the University of Chicago (Smith et al. 2021). For reasons elaborated below, although the cumulative data file contains 68,846 observations across 33 years of study, we

---

\*Extended.

will be focusing primarily on only respondents from the last two surveys, in 2018 and 2021. This significantly limits certain aspects of our investigation to 6,380 entries; however, the issues we highlight are endemic to the format and methodology of the survey itself, and thus do not require large magnitudes to afford predictive power.

In most years, the GSS is conducted in the form of a face-to-face interview, in which an interviewer poses a particular subset of questions in series to a respondent, and encodes their answer into a regularized format. To this end, the respondent is usually prompted with several options, e.g., “Agree,” “Neither agree nor disagree,” “Disagree,” from which they choose the most accurate response. In addition, some variables allow for ‘volunteered’ responses, which are codeable responses that are *not* read to the respondent, but rather only recorded if the respondent independently volunteers said response.

However, due to the constraints of the COVID-19 pandemic, the 2020 survey – which was postponed into 2021 – was partially conducted through an online survey platform (and the remainder via telephone). This platform did not offer volunteered responses, and respondents could only choose from the set of options presented to them. For 2021, the GSS bifurcated these variables in two, using either a -V or -NV suffix to indicate whether the respondent’s particular mode of communication afforded volunteered responses. As it happens, none of these variables are of study in this paper; however, as we are examining something so sensitive to wording as gender, it stands to reason that the absence of volunteered responses could have been detrimental to the granularity of description offered to respondents with which to communicate their identity. For example, the variable `SEXNOW1` – the 1 suffix indicating this specific phrasing of the question was *only* asked during the pandemic – only offers four options: “Male,” “Female,” “Transgender,” or “None of these.” This does not form a complementary set: someone identifying, say, as a transgender woman would be required to respond either “Female” or “Transgender,” but not both. If “transgender woman” had been available as a volunteered response, it is possible that some of the issues caused by this encoding would have been avoided.

The official GSS Codebook (Davern et al. 2021) indicates that the variables of the survey consist of a combination of respondent-offered responses (i.e., respondents’ answers to a given question) and interviewer-coded responses (i.e., questions that are not directly asked of the respondent, but rather recorded *about* the respondent, at the interviewer’s discretion). Again, for something as personal as gender or sexual identity, this necessarily introduces a cognitive gap that is subject to the interviewer’s personal perceptions and biases. As a pathological example, a particularly malicious interviewer could knowingly and willingly mis-code a transgender respondent’s sex (under the `SEX` variable) as their birth sex, rather than their chosen gender. As a result, we chose to de-emphasize `SEX` — an interviewer-coded variable asked across every ballot of every GSS since 1972 — in our study, in favour of the respondent-focused `SEXNOW` and `SEXNOW1`. These variables are only asked in a single year each – 2018 and 2021, respectively – which *significantly* reduces our pool of respondents. Future studies may be interested in a longitudinal view of how these new, more flexible questions of identity shape responses over time.

Finally, an important note on the usage of `NA` in the dataset: although the Codebook includes marginals for reserved non-response codes – “Don’t know,” “No answer,” “Skipped on web” – the software used to compile the data into an R dataframe (Healy 2019) condense all of these into the singular value of `NA`.

## Variables

Although 2021-specific variables like `SEXNOW1` are necessarily only answered by a small fraction of respondents, they use the same coding as their pre-2021 equivalents, and exist in complementary distribution to them (i.e., no respondent has a response to both variables). As a result, we were able to expand our selection by coalescing the variables together into a single column. This is notated below with a (1) suffix, although the 1-suffixed column was dropped in the actual dataframe. With this in mind, our summative list of variables is as follows:

- **YEAR**: The survey year in which the respondent was interviewed.
- **SEX**: The **interviewer-coded** sex of the respondent. (Used only for statistics on reporting, not for respondent identity.) Options: 1 (Male), 2 (Female).
- **SEXBIRTH(1)**: The assigned sex of the respondent at birth, based on physiognomy. Options: 1 (Male), 2 (Female), 3 (Intersex).
- **SEXNOW(1)**: The personally-identified sex/gender of the respondent, at survey time. Options: 1 (Male), 2 (Female), 3 (Transgender), 4 (None of these).
- **SEXORNT**: The sexual<sup>1</sup> orientation of the respondent. Options: 1 (Gay/Lesbian/Homosexual), 2 (Bisexual), 3 (Straight/Heterosexual).
- **SEXSEX**: The gender makeup of the respondent’s sexual partners, to their knowledge, for the past 12 months. Options: 1 (Exclusively male), 2 (Both male and female), 3 (Exclusively female).
- **SEXSEX5**: Similar to **SEXSEX**, but expanded over the past 5 years. Options: 1 (Exclusively male), 2 (Both male and female), 3 (Exclusively female).
- **PTNRORNT**: The orientation of the respondent’s *latest* sexual partner, to their knowledge. Options: 1 (Gay/Lesbian/Homosexual), 2 (Bisexual), 3 (Straight/Heterosexual), 4 (I have never had a sexual partner)<sup>2</sup>.

## Computed Variables

In addition to the above data variables, several additional columns were computed to make the associated data more portable and readable at-a-glance.

- **id**: A serial ID. Not used directly, but it was discovered that the serial IDs for respondents rolledback to 1 every survey year. For reproducibility, it was important to ensure that there were no ID collisions. So, the serial ID is now also prefixed with the year.
- **sex\_ornt\_label**: A compact, human-readable description of a person’s gender and sexual orientation, as might be used in everyday conversation. e.g., “Straight woman,” “Gay man,” “Bisexual trans person.”<sup>3</sup>

---

<sup>1</sup>There is an important distinction to be made between sexual attraction and romantic attraction, not to mention sexual attraction valency (i.e., the asexuality spectrum). However, for the purposes of this study and supplementary survey, only the prior is considered.

<sup>2</sup>This is the only variable under investigation which includes an option for “I have never had a sexual partner.” As a result, most analyses involving this variable will explicitly filter out this option, to keep a level playing field among the other sex-partner-related questions.

<sup>3</sup>Note that “gay” is used to refer to any kind of exclusively-homosexual attraction, as identified by the respondent. Historically, “gay” often referred exclusively to male homosexual attraction, with “lesbian” being more common for females; conversely, “gay” is currently entering common parlance for *any* same-gender attraction, even among those who also experience different-gender attraction. The usage of terms such as these to refer to different intersections of gender and sexuality over time is fascinating, and worthy of its own

## Methods

All data analysis<sup>4</sup> was performed in the R statistical programming language (R Core Team 2021), using RStudio (RStudio Team 2022) for ease of exploration, development, and production of this paper via built-in `knitr` (Xie 2021) functionality. GSS data are fetched by the `gssr` package (Healy 2019), which automatically packages the columns of the data into a ready-to-use dataframe. As usual, the `tidyverse` (Wickham et al. 2019) collection provides an indispensable array of resources, including but not limited to `ggplot2` (Wickham 2016) for visualization. For additional aesthetics, `kableExtra` (Zhu 2021), which<sup>5</sup> styles tables, and `patchwork` (Pedersen 2020), which composes multiple `ggplot` plots together, were also used.

## Results

### Respondent demographics

We first sought to develop an understanding of the demography of our respondents. In Figure 1 below, we see that respondents to the 2018 and 2021 GSS overwhelmingly self-identify as men and women (as opposed to transgendered or another gender), and predominantly as heterosexual. Interestingly, although the bisexual-identifying segment of the 2018 female respondent pool is so small as to be almost invisible, a significantly greater magnitude of them are present in the 2021 pool.

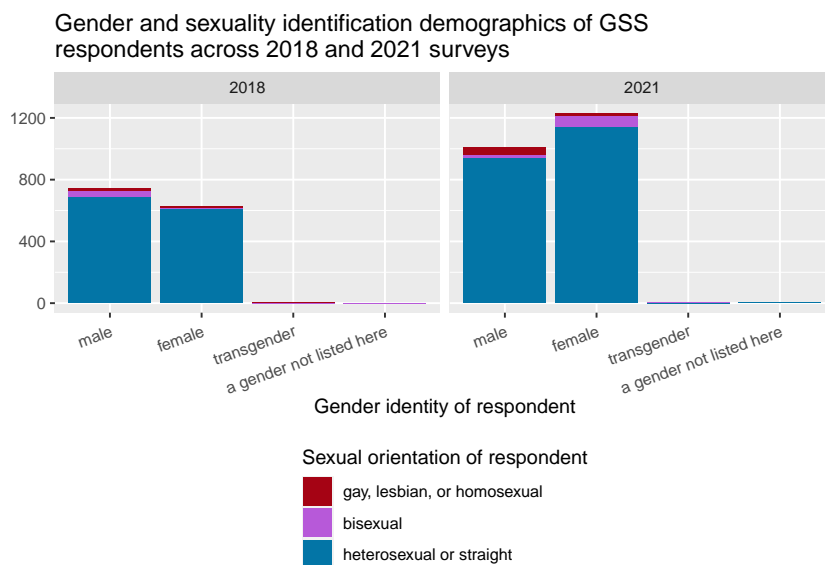


Figure 1: In both 2018 and 2021, respondent demographics are dominated by straight men and women. A significantly larger contingent of bisexual women responded in 2021.

study; however, it is immaterial to this study, and so “gay” was chosen as a cover-all term.

<sup>4</sup>Code available at [https://github.com/oliver-daniel/inf313\\_paper\\_3](https://github.com/oliver-daniel/inf313_paper_3).

<sup>5</sup>Attempts to, anyway. Half of the options cause Kable to dump raw  $\text{\LaTeX}$  and destroy the formatting of the entire document. Did you know that it’s literally impossible to move table captions to the bottom of the table?

Figure 2 zooms in on the non-binary gender options which, although representing an extremely small portion of the population – with no cohort representing more than ten respondents in a given year – exhibits a greater diversity in sexuality than their binary counterparts.

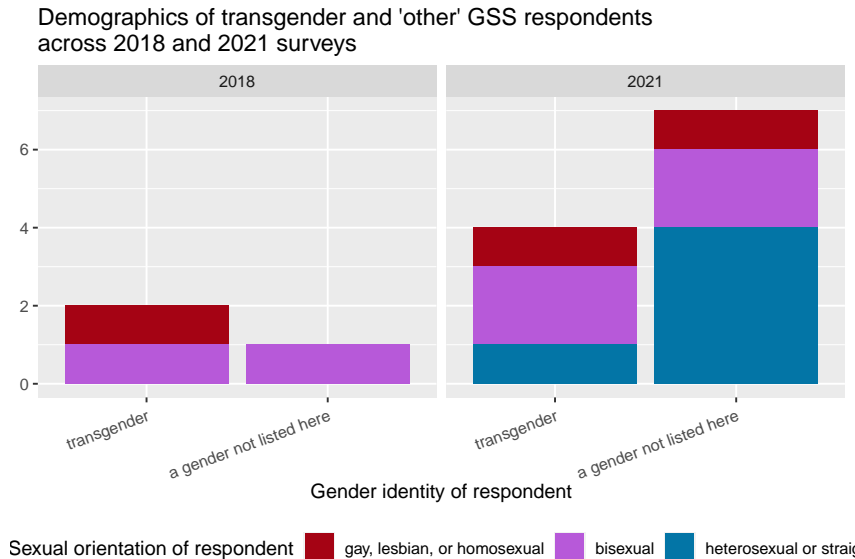


Figure 2: The overall turnout of transgender and 'other' (e.g., non-binary) respondents is low, but exhibits a wider diversity in sexuality.

With access to not only respondents' present gender identity, but also their birth sex, we were interested to determine the distinction, if any, between those respondents who responded 'transgender' as opposed to a binary gender. Table 1 contrasts birth sex (rows) with gender identity (columns) and counts total respondents at each intersection. As expected, the majority of respondents identify as cisgendered: that is, that their present gender corresponds to their birth sex. As many male-assigned respondents identified as female (i.e., as trans women) as transgender, implying that there is not a clear semantic delineation between the two. And, with the presence of a fourth 'other' category – which includes two male-assigned respondents – it is not necessarily the case that 'transgender' also includes non-binary and gender-non-conforming identities. This linguistic distinction is a clear area for future study. Fascinatingly, despite making up an estimated [TODO: amount, cite] of the U.S. population, only one respondent intimated that they were born intersex. As a result, and lacking access to this individual's actual interview/questionnaire, we will focus less on the complicated realm of post-intersex gender identity, especially as it corresponds to post-natal gender reassignment by medical practitioners; this, too, is a viable subject of intense study. [TODO CITE?]

### Interviewer gender coding and the SEX variable

As mentioned previously, early explorations of the data necessitated disregarding the **SEX** variable almost entirely. This is perhaps most strikingly exhibited below in Table 2; for brevity, the first column delineates the birth sex (M = male; F = female; I = intersex) and

Table 1: As expected, cisgendered respondents vastly outnumber any other configuration; in this dataset, the entire intersex population of the United States is represented by a single respondent.

	Male	Female	Transgender	Other
Assigned male	2472	6	6	2
Assigned female	8	2801	6	13
Intersex	0	1	0	0

the gender identification (M = male; F = female; T = transgender; X = other) with a slash.

As shown, more than 22% of self-reported cisgender women are recorded by interviewers as male, and 30% of self-reported cisgender men as female. Without further access to the specific methodologies of conducting a US-GSS interview, it is unknown to us what could have caused such a significant anomaly. The Codebook (Davern et al. 2021) notes:

The variable SEX was revised for the web mode as well. In face-to-face interviews, interviewers traditionally coded SEX based on interviewer observation. In the web mode, SEX was asked explicitly of respondents. Sex and gender identity are collected as two separate items: sex recorded at birth (SEXBIRTH1) and current gender identity (SEXNOW1). In the past, these items have only been asked to two-thirds of respondents (i.e., two ballots), on a self-administered topical module. Beginning with the 2020/2021 GSS (both the 2016-2020 GSS Panel Wave 2 and the 2021 GSS Cross-section), SEXBIRTH1 and SEXNOW1 are asked of all respondents. For the purpose of backward compatibility, the cross-section dataset contains SEX, which uses the traditional binary coding scheme, and is based on recoding both SEXBIRTH1 and SEXNOW1 in 2021 (which was collected by interviewer observation in prior years).

It is not at all clear what is meant by ‘recoding,’ in terms of how the SEX column is calculated from the aforementioned respondent-identified columns, especially when this calculation produced more than one-fifths’ error in what should have been the two most straightforward cases. It furthermore reflects poorly on the NORC that in each non-cisgendered combination, at least half of respondents are coded as their birth sex, even when their chosen gender is

Table 2: This table exhibits the first major inconsistency in the data.

	Recorded as male	Recorded as female
F/F	631	2170
F/M	7	1
F/T	0	6
F/X	0	13
I/F	1	0
M/F	3	3
M/M	1723	749
M/T	5	1
M/X	1	1

explicitly binary. For example, 100% of female-assigned respondents who identified as an unlisted gender were recorded as female, as were 50% of male-assigned, female respondents. Although we hesitate to draw from this a malicious reading, this inconsistency in gender-recording procedure, across multiple intersections of sex and gender, is certainly cause for suspicion.

"In terms of their orientation, my latest sexual partner was..."

by personal identity, as a proportion of respondents

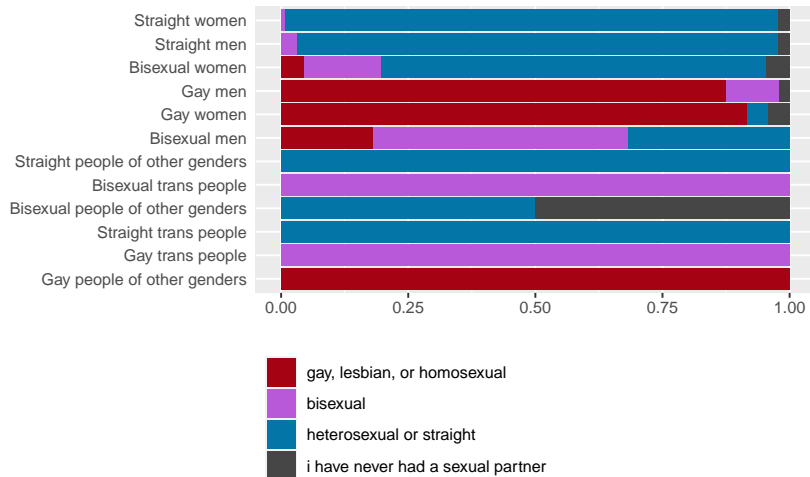


Figure 3: Surprisingly, the distributions of this variable seem to be more accurately reflective than those in Figure 4, in terms of what sorts of sexual partners one would expect a respondent to have had.

## Discussion

TODO

## Survey

TODO

"In terms of their sex, my sexual partners have been..."  
by personal identity, as a proportion of respondents

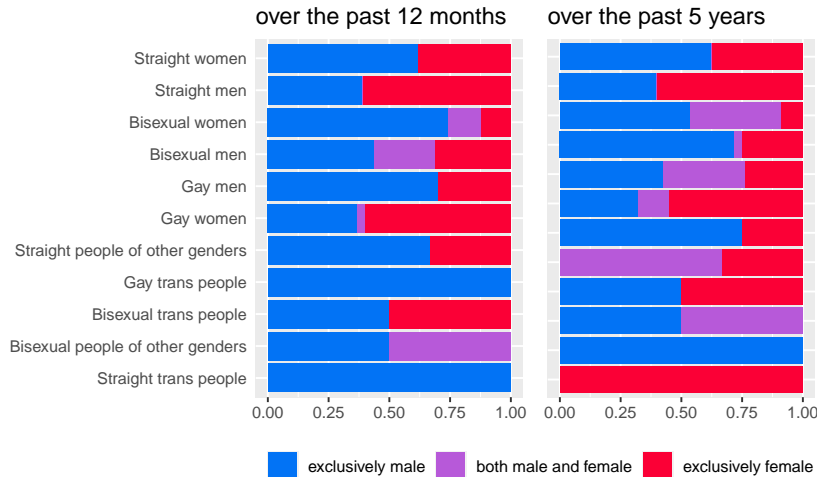


Figure 4: Although preference in sexual partners appears relatively stable over both 1-year and 5-year periods, an anomalously large contingent of heterosexual-identifying men and women have had exclusively same-sex relationships over these periods.

## References

- Davern, Michael, Rene Bautista, Jeremy Freese, and Stephen L. Morgan. 2021. *General Social Survey 2021 Cross-Section [Codebook]*.
- Healy, Kieran. 2019. *Gssr: General Social Survey Data for Use in r*. <http://kjhealy.github.io/gssr>.
- Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*. <https://CRAN.R-project.org/package=patchwork>.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- RStudio Team. 2022. *RStudio: Integrated Development Environment for r*. Boston, MA: RStudio, PBC. <http://www.rstudio.com/>.
- Smith, Tom W., Michael Davern, Jeremy Freese, and Stephen L. Morgan. 2021. *General Social Surveys, 1972-2021 [Machine-Readable Data File]*.
- Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolmund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. <https://yihui.org/knitr/>.
- Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. <https://CRAN.R-project.org/package=kableExtra>.