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

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[1] 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)[2].
- `PTNRSXNOW`: The gender identity of the respondent's *latest* sexual partner, to their knowledge. Options: 1 (Male), 2 (Female), 3 (Transgender), 4 (None of these).

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 respondent's gender and sexual orientation, as might be used in everyday conversation. e.g., "Straight woman," "Gay man," "Bisexual trans person."[3]

[1] 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.

[2] 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.

[3] 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

- `ptnr_label`: Similar, for respondents' latest sexual partner.

Methods

All data analysis[4] 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[5] 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 and proportion of them are present in the 2021 pool. The same trend is visible in the population of gay men, albeit to a lesser extent.

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.

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

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

[4] Code available at https://github.com/oliver-daniel/inf313_paper_3.

[5] Attempts to, anyway. Half of the options cause Kable to dump raw LaTeXand 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?

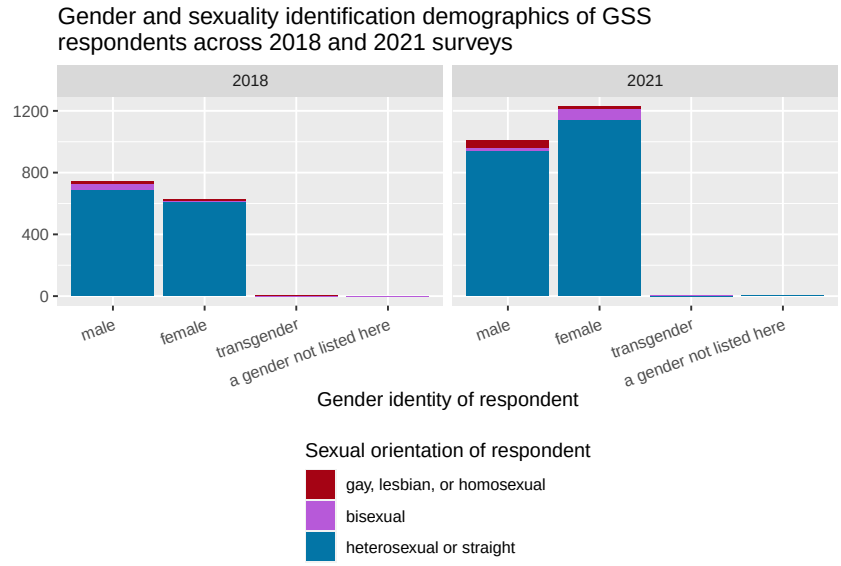

Figure 1: In both 2018 and 2021, respondent demographics are dominated by straight men and women. A significantly larger contingent of bisexual women responded in 2021.

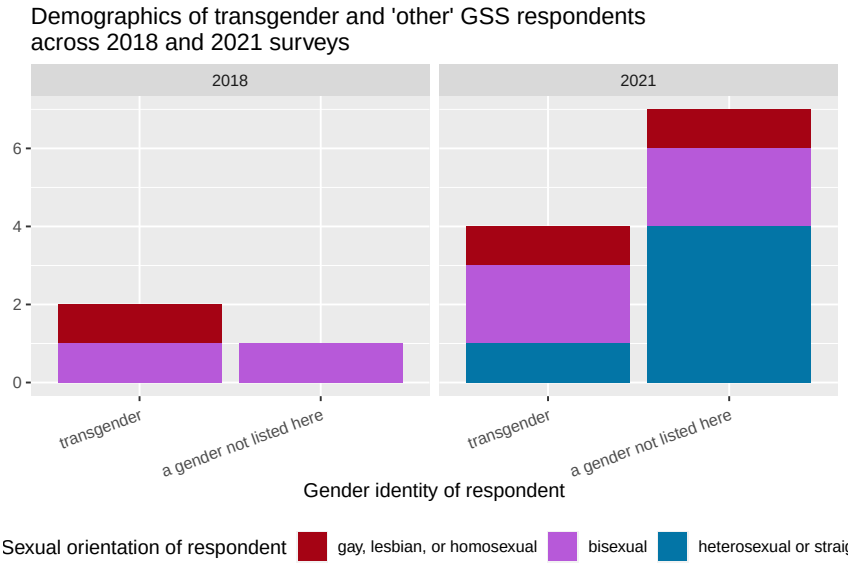

Figure 2: The overall turnout of transgender and 'other' (e.g., non-binary) respondents is low, but exhibits a wider diversity in sexuality.

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 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 explicitly different. 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.

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

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

Sexual orientation and recent sex partners

Tables 3-5 investigate the gender and sexuality makeup of respondents' recent sexual partners. The spans in question range from only the latest engagement, to a five-year period.

Figure 3 below provides a bird's-eye view of respondents' most recent sexual activity. Although a single sexual partner does not necessarily exemplify *all* of a respondent's partners over time, the recency of the data hopefully mitigates the risk of one's historical sexual behaviour being irreflective of their present orientation, especially in a Western heteronormative environment. This means, for example, that a person may have had several partners of the opposite sex before realizing that they were gay or otherwise interested in the same sex. As a result, their latest partner is likely – though by no means certain – to be someone to whose gender and sexuality the respondent is attracted.

Many high-frequency pairings are expected: straight men with straight women, gay men with gay men, and so on. Bisexual woman respondents in particular exhibit a wider variety in their partners than any other cohort of respondents, including some in which the partners' fields of attraction appear to mismatch. In one case, for example, a respondent who identified as a bisexual woman responded that their latest partner was a straight woman. In real life, such sexual pairings absolutely happen [TODO CITE? maybe some queer theory?] Similarly, a number of sexual encounters were recorded between straight men and straight men (5), and straight women with straight women (3). Surprisingly, transgender and non-binary (i.e., 'other') respondents exclusively reported binary-gendered partners. However, it must be acknowledged that, as such a narrow slice of the respondent pool, this absence could be by sheer chance.

Narrowing our focus to the sexuality (or existence) of the respondent's latest partner appears to confirm this: although binary-gendered respondents show at least some diversity in the proportions of their partners' sexualities, each row in Figure 4 below corresponding to a transgender or nonbinary cohort exhibits either 100% or 50% consensus, indicating very low magnitudes.

However, we run into another surprising anomaly when analyzing variables `SEXSEX` and `SEXSEX5`, which concern longitudinal sexual behaviour over 1- and 5-year periods, respectively. As might be expected, the 5-year results are usually slightly more diverse than their shorter-term counterparts; bisexual respondents, in particular, responded 'Both male and female'

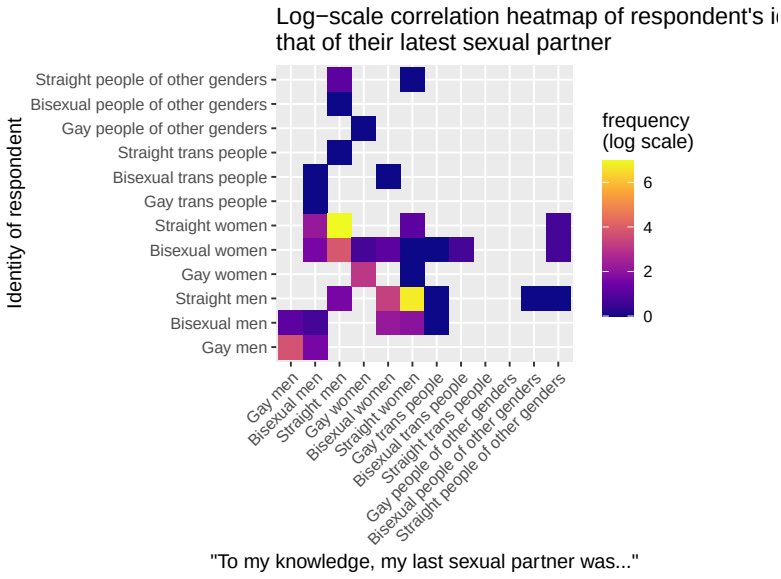

Figure 3: Including a log scale was necessary to compete with the overwhelmingly frequent pairings of straight men and straight women.

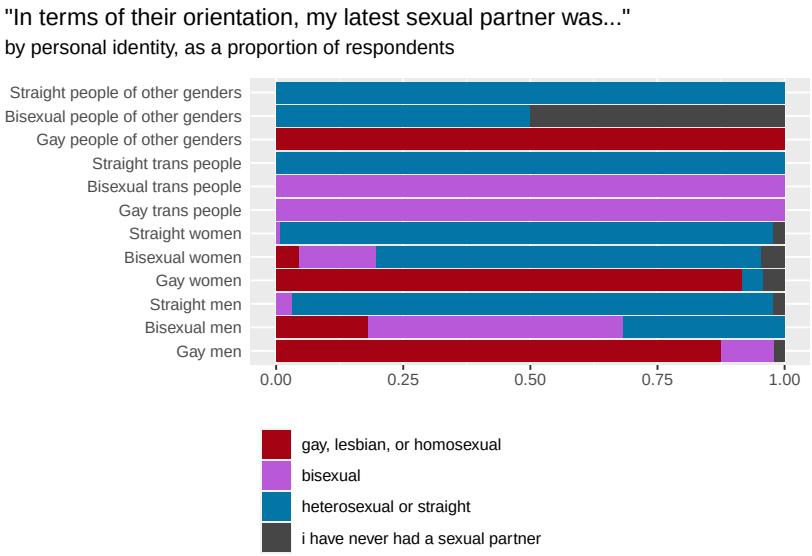

Figure 4: Surprisingly, the distributions of this variable seem to be more accurately reflective than those in Figure 5, in terms of what sorts of sexual partners one would expect a respondent to have had.

more frequently over the expanded timeframe. The sexual partners of straight men, on the other hand, appear hardly to change at all, save for a slightly-widened sliver of 'Both male and female' responses. That being said, one would imagine that these partners would be exclusively women – or that this would be the case for most straight men. Instead, almost 30% of straight male respondents report having had sex exclusively with men. Similarly, nearly a third of straight women report having sex only with women in the past year, a proportion that increases to full half when considering the past 5 years. Again, although we do not suppose to claim such a result is *impossible*, a 22-year-long observational study into the behaviour of straight people (Daniel et al, 2022)[6] appears to indicate this is at the least an *improbable* result. Instead, we propose that this is another notational error, similar to the miscoding of cisgendered respondents discussed above.

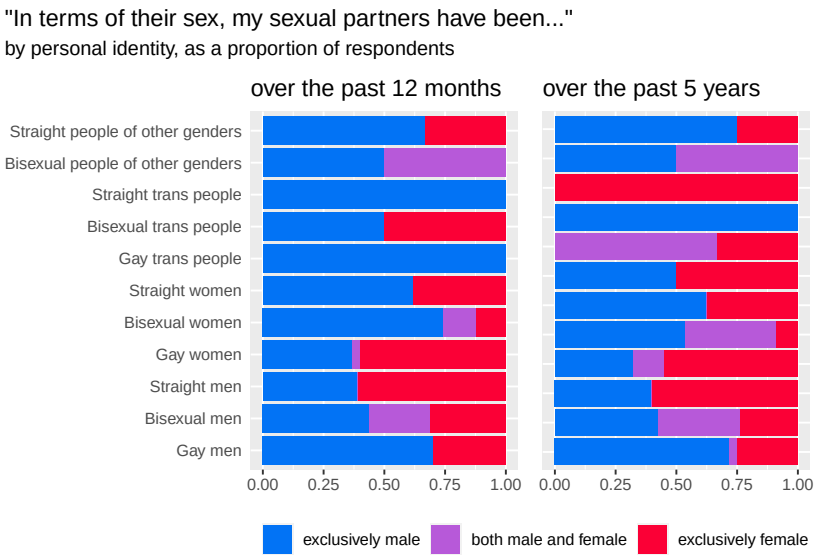

Figure 5: Although preference in sexual partners appears relatively stable over both 1-year and 5-year periods, an anomalously large contingent of heterosexual-identifying men and women have had exclusively same-sex relationships over these periods.

Table 3 investigates this anomaly further, specifically as it pertains to binary-gendered respondents with a single-sex attraction, e.g., respondents who identify as male or female, and gay or straight. In summary, it appears that having one's gender misrecorded in `SEX` but having had sexual partners exclusively of the expected sex, or vice versa, is significantly less likely than both variables being either expected or unexpected in value. To illustrate this confusing wording with an example, there are significantly more self-identified gay men who:

- a) were recorded as *men* and b) reported having sex with *only men* in the past year; or
- a) were recorded as *women* and b) reported having sex with *only women* in the past year

[6]Not cited; paper does not exist.

than some combination of the two. Mysteriously, gay women – the least abundant of the cohorts – reverse this trend, being the only cohort with any respondents who were misrecorded as men, but only having reported sex with women, and zero respondents to be accurately recorded as women with only female partners. This appears to be evidence of a comparatively widespread scrambling of responses when compared to personal identification, which may well have interfered with other findings in this paper. It is unknown at this time whether this confusion stems from survey respondents themselves, the interviewers who recorded their responses, at some point in the NORC datafiles, or even in the `gssr` client (Healy (2019)) we are using to fetch and process the data.

Discussion

TODO

Appendix: Survey

Our survey can be found online here.

Table 3: There appears to be a correlative effect between a mismatch in recorded sex, and the likelihood of sexual experiences beyond the expected demographic. Apart from that, there seems to be a confounding effect afoot.

	Sex miscoded?/Unexpected partner?			
Identity	No/No	No/Yes	Yes/Yes	Yes/No
Gay men	35	0	15	0
Straight men	739	5	471	0
Gay women	0	18	1	11
Straight women	774	8	474	0

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 Grolemund, 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.