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.

Introduction

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 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.

*Extended.

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.

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.)
- `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 respondents' 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).

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."[2]

Methods

All data analysis[3] 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.

Results

Discussion

Survey

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

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

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.

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/.