Title

An Automated Algorithm for Classifying Expansive Responses for Gender Identity

Authors

Ceja, Alexis
Raygani, Sawye
Conner, Bradley T
et al.

Publication Date

2023-11-07

Supplemental Material

https://escholarship.org/uc/item/5fx4d398#supplemental

An Automated Algorithm for Classifying Expansive Responses for Gender Identity

Ceja, Alexis, BA[1,2], Raygani, Sawye[2,3], Conner, Bradley T., PhD[4], Lisha, Nadra E., PhD[5], Bryant-Lees, Kinsey B., PhD[2,6], Lubensky, Micah E., PhD[1,2], Dastur, Zubin, MS, MPH[2,7], Lunn, Mitchell R., MD, MAS[2,8,9], Obedin-Maliver, Juno, MD, MPH, MAS[2,7,9], & Flentje, Annesa, PhD[1,2,10]

[1]Department of Community Health Systems, School of Nursing, University of California, San Francisco, San Francisco, CA, USA
[2]The PRIDE Study/PRIDEnet, Stanford University School of Medicine, Stanford, CA, USA
[3]Program in Human Biology, Stanford University Department of Humanities and Sciences, Stanford, CA, USA
[4]Department of Psychology, Colorado State University, Fort Collins, CO, USA
[5]Center for Tobacco Control, Research, and Education, Division of General Internal Medicine, Department of Medicine, University of California San Francisco, San Francisco, CA, USA
[6]Department of Psychological Sciences, Northern Kentucky University, Highland Heights, KY, USA
[7]Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford, CA, USA
[8]Division of Nephrology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA
[9]Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, USA
[10]Alliance Health Project, Department of Psychiatry and Behavioral Sciences, School of Medicine, University of California San Francisco, San Francisco, CA, USA

Corresponding author: Annesa Flentje, 2 Koret Way, N505, San Francisco, CA 94143, annesa.flentje@ucsf.edu, 415-502-0697

Disclosures: The authors declare the following relationships which may be considered potential conflicts of interest. Dr. Juno Obedin-Maliver has consulted for Sage Therapeutics (2017), Ibis Reproductive Health (2017-2018, 2020-present), Hims, Incorporated (2019-present), and Folx, Incorporated (2019-present). None of these engagements influenced or are pertinent to the work described in this manuscript. Dr. Mitchell R. Lunn has received consulting fees from Folx Health (2019 – 2021), Hims, Inc. (2019 – present), and Otsuka Pharmaceutical Development and Commercialization, Inc. (2023). The remaining authors have no conflicts of interest to disclose.

Funding: This project was funded by the National Institute on Drug Abuse and the Office of Research on Women's Health (R01DA052016 to Dr. Flentje). It was also supported by the National Institute on Drug Abuse (K23DA0398001 to Dr. Flentje), the National Institute of Diabetes and Digestive and Kidney Diseases (K12DK11028 to Dr. Obedin-Maliver), and the Patient-Centered Outcomes Research Institute® (PCORI®) (PPRN-1501-26848 to Dr. Lunn). The funding institutions did not contribute to the study design, collection and analysis of data, and writing of the manuscript.

Abstract

Objective: Current two-step measures of gender identity do not prescribe methods for using expanded responses (*i.e.*, multiple selections and open-ended responses) among sexual and gender minority (SGM) people, though SGM people want the opportunity to provide these responses. To increase the power of analyses using expanded gender identity responses, we created an automated algorithm to generate analyzable categories.

Materials and Methods: Participants' expanded gender identity responses and sex assigned at birth were used to create five categories (*i.e.*, cisgender men, cisgender women, gender expansive individuals, transgender men, and transgender women) from a cohort of SGM people (N = 6,312, 53% cisgender individuals). Data was collected from June 2020 to June 2021. Chi-square tests were performed.

Results: Forty-six percent of our sample may have been classified into an "other" category without an algorithm due to providing their own write-in response (5.7%), selecting "another gender identity" (5.7%), or selecting multiple (42.6%) or less commonly described (10.2%) gender identities. There was a relationship between the categories formed by our algorithm and participants' single category selection (χ2 [20] = 19,000, p < .001). Concordance rates were highest among participants classified by our algorithm as cisgender men, cisgender women, transgender men, and transgender women (percentages ranged from 97-99%), and lowest among participants classified as gender expansive individuals (74.3%).

Discussion: Without an algorithm to incorporate expanded gender identity responses, almost half of the sample may have been classified into an "other" category.

Conclusion: Our algorithm successfully classified participants into analyzable categories from expanded gender identity responses.

Keywords: gender identity; algorithm; sexual and gender minorities; methodology

Public significance statement: The gender identity categories that participants were classified into by our algorithm exhibited a high level of agreement with the single category participants selected from reduced, predefined answer choices for gender identity. This study's findings suggest that the automated algorithm we developed can be used to accurately and effectively classify participants into concise gender identity categories using expanded responses for gender identity and sex assigned at birth.

Introduction

Existing measures assessing gender identity often fail to account for the diverse lived experiences of gender minority (GM) people (*i.e.,* individuals whose gender identity differs from that often associated with their sex assigned at birth [SAAB]) (Reisner et al., 2015), despite the importance of accounting for GM status when studying health outcomes (Lefevor et al., 2019; Newcomb et al., 2019). To prevent the conflation of gender identity and sex and to ensure accurate classification of GM people, organizations like the National Academies of Sciences, Engineering, and Medicine (*Measuring Sex, Gender Identity, and Sexual Orientation*, 2022) have taken a two-step approach in which participants are asked about their gender identity and SAAB separately (Lombardi & Banik, 2016; Reisner, Biello, et al., 2014; Reisner, Conron, et al., 2014; Tate et al., 2013; The Gender Identity in U.S. Surveillance (GenIUSS) Group, 2014). Measures for gender identity may be limited by the use of forced-choice questions, single-select options, and exclusion of open-ended answer choices. These limitations may result in the misclassification and in some instances, the misgendering of GM people, which can reduce engagement (Tate et al., 2013), compromise the study's validity (Bauer et al., 2017), and hinder the identification and elimination of health disparities among this population (Patterson et al., 2017).

Prior work has demonstrated that sexual and gender minority (SGM) participants want the opportunity to select multiple answer choices and provide a written description of their gender identity (Beischel et al., 2022; Suen et al., 2020; Vivienne et al., 2021). Using expanded options for gender identity can make analysis of group differences difficult, and there remains little guidance on how to create analyzable categories that are inclusive of participants' diverse gender identities. Emerging methods that have been designed to create gender identity categories from expanded options (*e.g.,* the "gendercoder" package in R (Beaudry et al., n.d.) and a hierarchial clustering algorithm (Callander et al., 2021)) have not examined the concordance of these computer-generated categories with the category participants would choose when presented with reduced, single-select options; it is unknown if these algorithms place participants into the gender category that best describes them. Alternatively, some scholars have created an "other" category consisting of participants who selected multiple or less commonly described gender identities, provided their own write-in response, or selected "another gender identity." These data may be removed entirely from analyses because of the heterogeneity of these identities (Ridolfo et al., 2012). In order to use expanded gender identity measures and responses as preferred by SGM people, researchers must have methods for using these data to accurately capture the diverse identities of this population.

Objective

The purpose of this study was to establish an automated algorithm to create analyzable categories (*i.e.,* cisgender men, cisgender women, gender expansive individuals, transgender men, and transgender women) from participants' multiple selections and open-ended responses for gender identity. We aimed to determine the percentage of participants who may have been classified into an "other" category or removed from analyses without a method like this. We sought to identify the concordance of the gender identity categories established by our algorithm with the single category that participants chose from reduced, predefined answer choices (*i.e.,* "cisgender man," "cisgender woman," "non-binary," "transgender man," "transgender woman," and "another gender identity"). We also sought to identify if there were demographic differences (*i.e.,* age, education, geographic region, income, race and ethnicity, and sexual orientation) in the performance of our algorithm between participants who were classified into a gender identity category that was concordant *versus* discordant with their single category response.

Materials and Methods

Participants

Participants were enrolled in The PRIDE Study, a national, longitudinal cohort of SGM people. To be eligible for The PRIDE Study, participants must identify as lesbian, gay, bisexual, transgender, queer, or another sexual and/or gender minority, be 18 years or older, live in the United States or its territories, and be comfortable reading and writing in English. Participants are recruited through PRIDEnet Community Partners consisting of SGM health centers and organizations, in-person events, advertisement on social media, and by word-of-mouth (Lunn et al., 2019). Participants were included in the current study if they provided information about their gender identity in The PRIDE Study's 2020

Annual Questionnaire. Data collection was from June 2020 to June 2021. The PRIDE Study was approved by the institutional review boards of the University of California, San Francisco, Stanford University, and the WIRB-Copernicus Group (WCG). WCG IRB approved a request for a waiver of consent signature and a partial waiver of authorization for use and disclosure of protected health information (PHI). The IRB determined that a waiver of the signature requirement on the authorization for use of PHI was needed and approved for this research; therefore, subjects were not required to sign an authorization for the use and disclosure of PHI.

Measures

We assessed gender identity using two methods as shown in Figure 1. The first method permitted multiple selections for gender identity as well as providing a written description of their gender identity. Participants were asked, "What is your current gender identity?" with 12 choices: "agender," "cisgender man," "cisgender woman," "genderqueer," "man," "non-binary," "questioning," "transgender man," "transgender woman," "Two-spirit," "woman," and "another gender identity." Participants who selected "another gender identity" could provide a description of their gender identity. The first method appeared at the beginning of the 2020 Annual Questionnaire with other demographic variables (*i.e.,* SAAB, intersex status, and sexual orientation). This section was followed by the second method that involved self-classification into a single category from reduced, predefined answer choices: "If you had to choose only one of the following terms, which best describes your current gender identity?" with 6 choices (*i.e.,* "cisgender man," "cisgender woman," "non-binary," "transgender man," "transgender woman," and "another gender identity"). Here the term "cisgender" was defined and an example was provided. With this method, participants who selected "another gender identity" could not provide a written description of their gender identity. SAAB was assessed by asking participants: "What was the sex assigned to you at birth, for example, on your original birth certificate?" with answer choices of "female" and "male."

Additional demographics included age, education, geographic region, income, race and ethnicity, and sexual orientation. Education level was measured by asking participants their highest education level with 10 answer choices offered (*i.e.,* "no schooling;" "nursery school to high school, no diploma;" "high school graduate or equivalent (*e.g.,* GED);" "trade/technical/vocational training;" "some college;" "2-year college degree;" "4-year college degree;" "Master's degree;" "Doctoral degree;" and "Professional degree (*e.g.,* M.D., J.D., M.B.A.)." Geographic region was determined from participants' self-reported ZIP codes and was categorized into 4 U.S. Census Regions: Midwest, Northeast, South, and West. To assess income level, participants were asked to report their individual earnings (in US Dollars) before taxes and deductions from all sources in the 2019 tax year. Eighteen answer choices in increments of $9,999 (ranging from $0 to $200,001+) were provided. Race and ethnicity was assessed by the item: "Which categories best describe you?" with 8 options: "American Indian or Alaska Native;" "Asian;" "Black, African American or African;" "Hispanic, Latino or Spanish;" "Middle Eastern or North African;" "Native Hawaiian or other Pacific Islander;" "White;" and "none of these fully describe me." Participants who selected "none of these fully describe me" could provide a written description of their race and/or ethnicity. Sexual orientation was assessed by the item, "What is your current sexual orientation?" with 11 options (*i.e.,* "asexual," "bisexual," "gay," "lesbian," "pansexual," "queer," "questioning," "same-gender loving," "straight/heterosexual," "Two-spirit," and "another sexual orientation" with an accompanying open-ended response). Multiple answer choices were permitted for race and ethnicity and for sexual orientation.

Automated algorithm

We developed an automated algorithm in which we classified participants into 5 mutually exclusive categories (*i.e.,* cisgender men, cisgender women, gender expansive individuals, transgender men, and transgender women) based on expanded responses for gender identity and SAAB. We used the terms "transgender men" and "transgender women" to capture identities existing within the gender binary, which may have included identities as "man" or "woman" but may have also included identities on the binary spectrum that were not "man" or "woman" (*e.g.,* "masculine"). We defined cisgender men as participants who reported one or more gender identities exclusively within the masculine binary (*e.g.,* "man" and "cisgender man") and were assigned male sex at birth. No participants classified through our algorithm as cisgender men provided a written description of their gender identity. We defined cisgender women as participants who reported one or more gender identities exclusively within the feminine binary (*e.g.,* "woman" and "cisgender woman") and were assigned female sex at birth. Examples of open-ended responses included "femme" and "lesbian." We defined gender expansive individuals as participants who reported one or more gender identities beyond the binary (*e.g.,* "agender," "genderqueer," "non-binary," "questioning," and "Two-spirit"). Examples of open-ended responses included "bigender," "genderfluid," and "neutrois." We defined transgender women as participants who reported one or more gender identities exclusively within the feminine binary (*e.g.,* "transgender woman" and "woman") and were assigned male sex at birth. An example of an open-ended response was "transsexual woman." We defined transgender men as participants who reported one or more gender identities exclusively within the masculine binary (*e.g.,* "transgender man" and "man") and were assigned female sex at birth. Examples of open-ended responses included "man of trans experience" and "transmasculine."

Our algorithm used open-ended responses to identify strings associated with five terms related to gender identity – cisgender, feminine, masculine, non-binary, and transgender – and incorporated the coding of these strings into the final category determination. Examples of strings associated with these terms are available through Figure 2. We utilized open-ended responses to identify strings associated with intersex status (*e.g.,* "intersex"); however, this coding only determined the final category assigned for select cases. For instance, participants who wrote an intersex-related identity and selected no other gender identity responses were classified as gender expansive individuals. Our algorithm was developed and refined using data from The PRIDE Study's 2017-2020 Annual Questionnaires. Our algorithm is provided in Supplementary Materials 1 (PDF version of the code), 2 (Stata version of the code) and 3 (R version of the code). Examples of gender identity classification using actual participant responses is presented in Table 1.

Analysis

Data were analyzed using Stata software, version 17 (*StataCorp, 2021*). We performed a chi-square test of independence to examine whether there were demographic differences (*i.e.,* age, education, geographic region, income, race/ethnicity, and sexual orientation) between the gender identity categories participants were assigned to using our algorithm. We tabulated the frequency of participants who may have been classified into an “other” category or dropped from analyses without an algorithm due to selecting “another gender identity,” providing their own write-in response, or selecting multiple or less commonly described gender identities (*e.g.,* agender). To pursue the primary aims of the study, we conducted separate sets of chi-square (χ^2) tests of independence. First, we examined the relationship between the two methods of assessing gender identity (*i.e.,* using our algorithm and the single category participants chose from reduced, predefined answer choices). Second, we examined whether there were demographic differences (*i.e.,* age, education, geographic region, income, race and ethnicity, and sexual orientation) between participants who were classified into a gender identity category by our algorithm corresponding with their single category selection (classification concordant) and those who were placed in a category that did not correspond with their single category selection (classification discordant). Concordance for the gender expansive category was met when participants selected “non-binary” or “another gender identity” as their single category response. For analysis of differences by age, we created four categories: (1) 18-30, (2) 31-45, (3) 46-60, and (4) 61 and older (Mehta et al., 2020). Since the categories for the race and ethnicity, and sexual orientation measures were not mutually exclusive, we created dichotomous variables for each category and assessed each in separate tests. Given the large sample size and number of tests, alpha was set to .01 to reduce the probability of Type II error (Westgard & Groth, 1979).

Data availability

Due to ethical restrictions related to sensitive participant information, study data can be made available on request in accordance with certain data access conditions by contacting research@pridestudy.org.

Code availability

The algorithm to classify expanded responses for gender identity is available within the supplementary materials.

Results

Of the 6,312 participants in the sample, 23.3% (*n* = 1,473) were classified as cisgender men, 30.1% (*n* = 1,898) as cisgender women, 33% (*n* = 2,086) as gender expansive individuals, 9.3% (*n* = 586) as transgender men, and 4.3% (*n* = 269) as transgender women. The majority of the sample identified as White (90%), had a college, graduate, or professional degree (76.4%), and lived in the Western or Southern regions of the U.S. (59.2%). Participants classified as gender expansive and transgender men had lower median ages (27.2 and 27.5, respectively) than those classified as cisgender men (42.9), cisgender women (30.9), and transgender women (41.6). Participants classified as cisgender men, and transgender women often reported an income above $40,000, whereas participants classified as cisgender women, gender expansive and transgender men typically reported an income below $40,000. The most frequently endorsed sexual orientation among participants classified as cisgender men was gay. The sexual orientations of lesbian and queer were often reported by participants classified as cisgender women. Participants who were classified as gender expansive typically identified as queer, participants who were classified as transgender men commonly identified as bisexual and queer, and participants who were classified as transgender women frequently identified as lesbian.

Approximately 1.3% (*n* = 81) of our sample did not select a category from the reduced, predefined answer choice list, yet provided information about their gender identity when given expanded options. Among these participants, 19.8% (*n* = 16) were classified by our algorithm as cisgender men, 33.3% (*n* =27) as cisgender women, 27.2% (*n* = 22) as gender expansive, 16% (*n* = 13) as transgender men, and 3.7% (*n* = 3) as transgender women.

Demographic differences between individuals in categories created through our algorithm

Differences in demographics by gender identity category created through our algorithm are presented in Table 2. The individuals in categories formed through our algorithm significantly differed by age category (χ^2 [12] = 1,000, $p < .001$), education (χ^2 [12] = 299.6, $p < .001$), geographic region (χ^2 [12] = 83.6, $p < .001$), income (χ^2 [12] = 597.2, $p < .001$), and sexual orientation (χ^2 [4] = 22.9-2,900, $p < .001$ for all). In addition, there were differences by specific racial and ethnic categories including American

Indian or Alaska Native (χ^2 [4] = 13.3, p = .010); Middle Eastern or North African (χ^2 [4] = 13.8, p = .008); White (χ^2 [4] = 15.9, p = .003); participants who reported that "none of the [racial and ethnic] categories fully described [them]" (χ^2 [4] = 20.1, p < .001), and participants who selected more than one race and/or ethnicity (χ^2 [4] = 42.3, p < .001).

Participants who may have been classified into an "other" category or removed from analyses without an algorithm

Almost half of our sample (44.6%, n = 2,813) may have been classified into an "other" category or dropped from analyses without an algorithm including participants who selected "another gender identity" (5.7%, n = 362), provided their own write-in response (5.7%, n = 359), or reported multiple (42.6%, n = 2,687) or less commonly described (*i.e.,* agender, questioning, and Two-spirit, 10.2%, n = 646) gender identities.

Two methods for assessing gender identity

The gender identity categories created through our algorithm and the single category participants chose from reduced, predefined answer choices were significantly related (χ^2 [20] = 19,000, p < .001). The observed frequencies are reported in Table 2. Visual inspection of the χ^2 table indicated that our algorithm successfully assigned the same category that participants chose for their single category selection under the reduced choice model in at least 97% of the cases except for gender expansive individuals (74.3% match): cisgender man (98.8% match), cisgender women (99.0% match), transgender men (98.1% match), and transgender women (97.0% match).

We identified several common response patterns among participants who were classified into a gender identity category by our algorithm that was discordant with their single category selection under the reduced choice model. Among the participants classified as cisgender men through our algorithm who self-selected a category other than "cisgender man" (n = 17), 82.4% identified as a "man" under the expanded choice model and chose "another gender identity" (n =12) or "non-binary" (n = 2) as their single category response under the reduced choice model. A similar pattern emerged for participants who were classified as cisgender women through our algorithm but self-selected a category other than "cisgender woman" (*n* = 19); most (94.7%) identified as a "woman" under the expanded choice model and chose "another gender identity" (*n* =14) or "non-binary" (*n* = 4) as their single category selection under the reduced choice model. Participants who were classified by our algorithm as gender expansive and selected a category other than "another gender identity" or "non-binary" under the reduced choice model (*n* = 530) typically identified with one or more gender identities beyond the binary (*e.g.,* "cisgender woman" and "non-binary") under the expanded choice model. Among these participants, 12.8% (*n* = 68) self-selected "cisgender man," 40.6% (*n* = 215) "cisgender woman," 35.9% (*n* = 190) "transgender men," and 10.8% (*n* = 57) "transgender women" as their single category response under the reduced choice model. The majority of these participants endorsed one or more gender expansive term (*i.e.,* "agender," "genderqueer," "non-binary," "questioning," and "Two-spirit") (*n* = 507); this may have been in addition to other gender identity answer choices. Of the participants categorized as transgender men through our algorithm who chose a category other than "transgender man" (*n* = 11), most identified with binary terms that did not include the word *cisgender* (*i.e.,* "transgender man" and "man") under the expanded choice model and chose "another gender identity" (*n* = 4), "cisgender man" (*n* = 4), or "non-binary" (*n* = 2) as their single category response. Similarly, participants categorized as transgender women through our algorithm who chose a category other than "transgender woman" (*n* = 8) typically identified as a "woman" and/or "transgender woman" under the expanded choice model and selected "another gender identity" (*n* = 4), "cisgender woman" (*n* = 3), or "non-binary" (*n* = 1) as their single category response.

Demographic differences between participants in concordant *versus* discordant categories

Demographic information of participants who were classified in a gender identity category concordant *versus* discordant with the single category they selected from reduced, predefined answer choices is presented in Table 3. These two groups differed significantly by age category ($p < .001$), education ($p < .001$), and income ($p < .001$). There were significant differences between these groups in the endorsement of specific racial or ethnic and sexual orientation categories including American Indian or Alaska Native (p = .014), participants who reported that "none of the [racial and ethnic] categories fully described [them]" (p = .007), bisexual (p < .001), gay (p = .001), pansexual (p < .001), queer (p < .001), questioning (p < .001), same-gender loving (p < .001), Two-spirit (p = .013), and participants who reported more than one sexual orientation (p < .001). Geographic region and the remaining racial and ethnic and sexual orientation categories were not associated with significant differences between participants who were classified by our algorithm into a category concordant *versus* discordant with their single reduced, predefined category response (p = .211-.898).

Discussion

Despite a growing body of research illustrating the desire from SGM communities for expanded gender identity choices (Beischel et al., 2022; Suen et al., 2020; Vivienne et al., 2021), this is the first study, to our knowledge, that has established an automated algorithm to create analyzable categories to account for these expanded choices among a large, diverse sample of SGM participants and compared these categories with participants' single reduced category selection. The gender identity categories established through our algorithm differed significantly by age category, education, geographic region, income, sexual orientation, and specific racial and ethnic categories including American Indian or Alaska Native; Middle Eastern or North African; White; participants who reported that "none of the [racial and ethnic] categories fully described [them]," and participants who selected more than one race and/or ethnicity. These findings challenge the idea that gender identity can be treated as a covariate in analyses because it is not equally distributed across demographic characteristics. Methods will be needed to identify how to incorporate gender identity into future studies without assuming an equal distribution of gender identity.

Consistent with prior work (Kuper et al., 2012; Lunn et al., 2019), a large proportion (44.6%) of our sample selected multiple or less common gender identities, provided their own write-in response, or selected "another gender identity" underscoring the importance of including expanded choices in measures of gender identity. These participants may have been grouped into an "other" category or dropped from analyses completely. Our findings indicated that the gender identity categories created through our algorithm were highly related to their single category selection. Accurate classification of gender identity into more concise categories is important to increase the statistical power of the analyses for group comparison. By having sufficient statistical power, analyses can improve knowledge about and identification of SGM subgroups at greater risk for specific health outcomes and contribute to effective resource allocation.

When examining the relationship between the two methods of assessing gender identity (*i.e.,* our algorithm and the single category participants chose from reduced, predefined answer choices), our algorithm was least effective for participants who were classified by our algorithm as gender expansive. A quarter (25.7%) of participants who were classified as gender expansive individuals by our algorithm did not select "non-binary" or "another gender identity" as their single category selection. These participants selected terms beyond the binary *(e.g.,* non-binary) when offered expanded options, but most chose a binary gender (*e.g.,* cisgender woman) for their single category selection. There are both pros and cons to researchers determining this classification versus participants self-selecting their own gender category. For instance, if researchers determine a gender grouping, they can ensure that participants are grouped by similar responses for gender identity and SAAB. When participants self-select from reduced, predefined answer choices, some participants may be grouped together that differ significantly in their lived experiences and identities due to differences in the understanding and usage of certain terms (*e.g.,* cisgender), or decision making processes while determining which category to select. On the other hand, when researchers use a reduced category that the participant has chosen, it gives the participant more autonomy and decision-making power in how they are grouped within research. More research is needed to understand if participants who were classified as gender expansive through our algorithm whose classification was discordant with their single category selection have health outcomes more similar to those who were classified as gender expansive and selected "another gender identity" or "non-binary" for their single category selection, or to participants who were classified into a binary category (*e.g.,* cisgender women). Our algorithm may reveal underlying differences in gender identity that are not captured when participants are required to select one category from a reduced, predefined answer choice list. Alternatively, participants' selection of a single gender identity category that best describes them may be a better metric of related health outcomes than a category chosen by researchers.

The algorithm worked well for participants in the remaining four gender identity categories: cisgender men, cisgender women, transgender men, and transgender women (concordance rates ranged from 97-99%). Nearly all participants who were classified by our algorithm as cisgender men or women self-selected "cisgender man" or "cisgender woman" for their single category selection, yet a small percentage self-selected "another gender identity." This may be due to a dislike of the term "cisgender" as utilized in "cisgender men" and "cisgender women," which were two options from the reduced, predefined answer choice list. Although we define the term "cisgender" in our annual questionnaires, we have received direct feedback from some older cisgender participants who are sexual minority (*e.g.,* cisgender lesbian women) about their dislike for the term because it does not reflect their lived experiences. This dislike may be attributed to generational differences in the language and conceptualization of gender identity and lack of transgender-inclusive measures in research with older sexual minority adults (Institute of Medicine, 2011).

Age category, income, and education significantly differed between participants who were classified by our algorithm into a category concordant *versus* discordant with their single category response. In addition, there were significant differences between concordance and discordance among participants who were American Indian or Alaska Native and endorsed that "none of the [racial and ethnic] categories fully described [them]." There were differences among participants who identified their sexual orientation as bisexual, gay, pansexual, queer, questioning, same-gender loving, Two-spirit, and participants who reported more than one sexual orientation. Future work is needed to explore potential reasons for these demographic differences. Replication of our algorithm in other data sets could provide insight into whether these findings are unique to the current study or evident across datasets.

Limitations

There were several important limitations to this work. Despite our large sample we had comparatively smaller samples of transgender men (9.3%) and transgender women (4.3%). The sample was predominantly White only (81.4%), though we did have large enough samples of racial or ethnic minority SGM people to enable comparisons by race and ethnicity. Implementation of the algorithm in samples with greater representation from transgender men and women, in samples with greater racial and ethnic diversity, and in non-SGM samples is imperative to ensure that our algorithm can capture the identities and experiences of these subpopulations and to identify new language and terminology used to describe gender identity. The replicability of our algorithm in other datasets is dependent upon the answer choices provided in the gender identity measure used. Studies using measures with reduced answer choices may have more diverse open-ended responses than those with a broader range of choices, which may require more time to modify the algorithm to recognize and correctly classify participants based on these additional responses.

Conclusion

Researchers are encouraged to assess gender identity using a two-step approach and provide participants with the option of selecting multiple responses and including their own write-in response. The obtained data can be difficult to analyze, however, through the implementation of our algorithm, we were able to successfully classify participants into concise gender identity categories; these classifications were highly concordant with their single category selection. This suggests that the algorithm may be used and adapted for other data sets to create analyzable categories from expanded answer choices for gender identity. Demographic differences were found between participants in the gender identity categories created by our algorithm and between participants who were classified into a category concordant *versus* discordant with their single category response. Without our algorithm, nearly half of our sample may not have been accurately classified into a gender identity category and may have been placed into an “other” category or removed from analyses. Rigorous methodology must be applied in research and clinical practice to ensure appropriate classification of gender identity for SGM people who have diverse identities and lived experiences.

Table 1

Example gender identity classification using actual participant responses

Gender Identity Category	Selection(s) using the Expanded Choice Model	Open-Ended Gender Identity Response	Participant-Reported Sex Assigned at Birth	Algorithm Number
Cisgender men	"Cisgender man"	None	"Male"	6
Cisgender woman	"Cisgender woman"	None	"Female"	1
	"Another gender identity" and "Cisgender woman" and "Woman"	"Femme"	"Female"	3
Gender expansive	"Agender" and "Non-binary"	None	"Female"	19
	"Cisgender man" and "Woman"	None	"Female"	20
Transgender men	"Transgender man"	None	"Female"	11
	"Another gender identity" and "Man" and "Transgender man"	"Transmasculine"	"Female"	12
Transgender women	"Transgender woman"	None	"Male"	15
	"Another gender identity" and "Transgender woman," and "Woman"	"Lesbian"	None	17

Note. Please see Supplemental Materials 1, 2, or 3 for all algorithms.

Table 2

Demographics of participants who reported gender identity information by the category assigned through our algorithm in The PRIDE Study's 2020 Annual Questionnaire (N = 6,312)

Variable	Cisgender men (*N* = 1,473)	Cisgender women (*N* = 1,898)	Gender expansive individuals (*N* = 2,086)	Transgender men (*N* = 586)	Transgender women (*N* = 269)	*p*
Expanded gender identity answer choices[a] (*n,* %)						
Agender	0 (0)	0 (0)	318 (15.2)	0 (0)	0 (0)	
Cisgender man	905 (61.4)	0 (0)	47 (2.3)	3 (.5)	0 (0)	
Cisgender woman	0 (0)	1,487 (78.3)	138 (6.6)	0 (0)	2 (.7)	
Genderqueer	0 (0)	0 (0)	845 (40.5)	0 (0)	0 (0)	
Man	874 (59.3)	0 (0)	166 (8)	290 (49.5)	0 (0)	
Non-binary	0 (0)	0 (0)	1,465 (70.2)	0 (0)	0 (0)	
Questioning	0 (0)	0 (0)	295 (14.1)	0 (0)	0 (0)	
Transgender man	0 (0)	0 (0)	291 (14)	553 (94.3)	0 (0)	
Transgender woman	0 (0)	0 (0)	90 (4.3)	0 (0)	255 (94.8)	
Two-spirit	0 (0)	0 (0)	54 (2.6)	0 (0)	0 (0)	
Woman	0 (0)	891 (46.9)	378 (18.1)	0 (0)	126 (46.8)	
Another gender identity	0 (0)	14 (.7)	332 (15.9)	14 (2.4)	2 (.7)	
Selected more than one gender identity	306 (20.8)	490 (25.8)	1,507 (72.2)	269 (45.9)	115 (42.8)	
Reduced, single-select gender identity answer choices (*n,* %)						
Cisgender man	1,440 (98.8)	1 (.1)	68 (3.3)	4 (.7)	0 (0)	
Cisgender woman	1 (.1)	1,852 (99)	215 (10.4)	0 (0)	3 (1.1)	
Non-binary	2 (.1)	4 (.2)	1,362 (66)	2 (.4)	1 (.4)	
Transgender man	1 (.1)	0 (0)	190 (9.2)	562 (98.1)	0 (0)	
Transgender woman	1 (.1)	0 (0)	57 (2.8)	1 (.2)	258 (97)	
Another gender identity	12 (.8)	14 (.7)	172 (8.3)	4 (.7)	4 (1.5)	
Sex assigned at birth (*n,* %)						
Female	0 (0)	1,898 (100)	1,795 (86.3)	583 (99.7)	1 (.4)	

Male	1,473 (100)	0 (0)	286 (13.8)	2 (.3)	267 (99.6)	
Age, in years, median (IQR)	42.9 (30.9-57.5)	30.9 (25.3-40.3)	27.2 (23-33.7)	27.5 (21.8-35.8)	41.6 (30.5-56.8)	**< .001**
Race and ethnicity[a] (*n,* %)						
American Indian or Alaska Native	32 (2.2)	53 (2.8)	86 (4.1)	24 (4.1)	10 (3.7)	**.010**
Asian	71 (4.8)	89 (4.7)	131 (6.3)	27 (4.6)	12 (4.5)	.134
Black, African American, or African	57 (3.9)	73 (3.9)	93 (4.4)	27 (4.6)	8 (3.0)	.654
Hispanic, Latino, or Spanish	121 (8.2)	127 (6.7)	141 (6.8)	47 (8)	17 (6.3)	.340
Middle Eastern or North African	13 (.9)	22 (1.2)	45 (2.2)	5 (.9)	3 (1.1)	**.008**
Native Hawaiian or other Pacific Islander	1 (.1)	5 (.3)	5 (.2)	2 (.3)	1 (.4)	.661
White	1,288 (87.4)	1,733 (91.3)	1,880 (90.1)	530 (90.4)	248 (92.2)	**.003**
None of these fully describe me	12 (.8)	25 (1.3)	49 (2.3)	3 (.5)	6 (2.2)	**< .001**
Selected more than one race and/or ethnicity	111 (7.5)	205 (10.8)	300 (14.4)	76 (13)	29 (10.8)	**< .001**
Sexual orientation[a] (*n,* %)						
Asexual	19 (1.3)	153 (8.1)	412 (19.8)	60 (10.2)	34 (12.6)	**< .001**
Bisexual	166 (11.3)	757 (40)	736 (35.3)	199 (34)	80 (29.7)	**< .001**
Gay	1,320 (89.6)	234 (12.3)	344 (16.5)	178 (30.4)	10 (3.7)	**< .001**
Lesbian	0 (0)	898 (47.3)	385 (18.7)	8 (1.4)	140 (52)	**< .001**
Pansexual	46 (3.1)	302 (15.9)	483 (23.2)	99 (16.9)	48 (17.8)	**< .001**
Queer	176 (12)	794 (41.8)	1,351 (64.8)	261 (44.5)	61 (22.7)	**< .001**
Questioning	9 (.6)	27 (1.4)	78 (3.7)	33 (5.6)	19 (7.1)	**< .001**
Same-gender loving	46 (3.1)	80 (4.2)	133 (6.4)	30 (5.1)	10 (3.7)	**< .001**
Straight/heterosexual	4 (.3)	8 (.4)	24 (1.2)	61 (10.4)	17 (6.3)	**< .001**
Two-spirit	4 (.3)	2 (.1)	27 (1.3)	1 (.2)	4 (1.5)	**< .001**
Another sexual orientation	9 (.6)	50 (2.6)	146 (7)	18 (3.1)	12 (4.5)	**< .001**
Selected more than one sexual orientation	254 (17.2)	946 (49.8)	1,287 (61.7)	252 (43)	117 (43.5)	**< .001**
Income level (*n,* %)						**< .001**
≤ $20,000	229 (17.1)	501 (29.5)	891 (47.7)	263 (49.5)	79 (32)	
$20,001 to $40,000	267 (20)	385 (22.7)	437 (23.4)	112 (21.1)	45 (18.2)	

$40,001 to $60,000	211 (15.8)	287 (16.9)	268 (14.3)	65 (12.2)	36 (14.6)	
≥ $60,001	631 (47.2)	523 (30.8)	272 (14.6)	89 (16.8)	87 (35.2)	
Education level (*n,* %)						**< .001**
No high school diploma	2 (.1)	7 (.4)	16 (.9)	9 (1.7)	1 (.4)	
High school/GED graduate or some college	236 (17.5)	270 (15.8)	538 (28.6)	204 (38.1)	70 (28.1)	
College degree (2- or 4-year)	504 (37.4)	647 (37.9)	802 (42.6)	198 (37)	108 (43.4)	
Graduate or Professional degree	606 (45)	784 (45.9)	525 (27.9)	124 (23.2)	70 (28.1)	
Geographic region (*n,* %)						**< .001**
Midwest	247 (17)	379 (20.3)	459 (22.5)	117 (20.5)	48 (18.2)	
Northeast	236 (16.2)	427 (22.9)	450 (22)	121 (21.2)	45 (17.1)	
South	475 (32.7)	432 (23.2)	490 (24)	162 (28.3)	96 (36.4)	
West	497 (34.2)	625 (33.6)	643 (31.5)	172 (30.1)	75 (28.4)	

[a.] Multiple answer choices were allowed.

Table 3

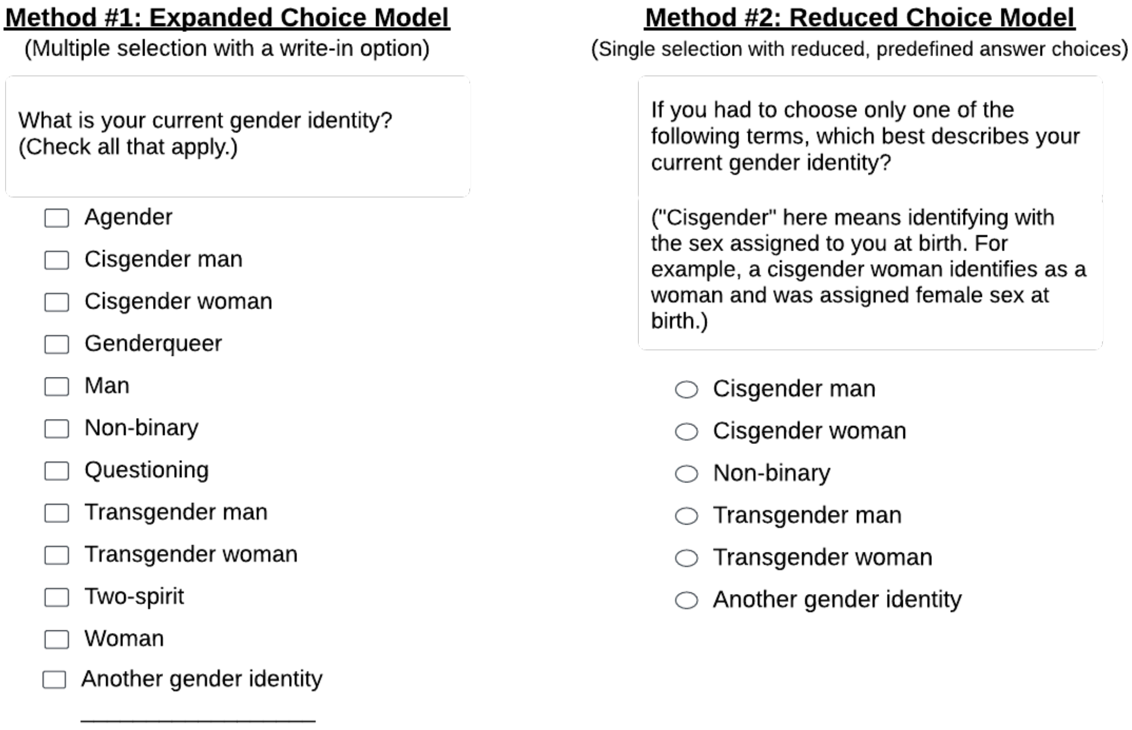
Demographic differences between participants who were classified by our algorithm into a gender identity category that was concordant versus discordant with their single category selection

	Classification concordant	Classification discordant	
	(*N* = 5,646)	(*N* = 585)	
Variable	(*n*, %)	(*n*, %)	*p*
Age category			**< .001**
18-30 years	2,771 (89.2)	335 (10.8)	
31-45 years	1,623 (91.5)	151 (8.5)	
46-60 years	790 (94.3)	48 (5.7)	
61 years and older	462 (90.1)	51 (9.9)	
Race and ethnicity[a]			
American Indian or Alaska Native	173 (85.6)	29 (14.4)	**.014**
Asian	298 (91.1)	29 (8.9)	.740
Black, African American, or African	225 (88.9)	28 (11.1)	.350
Hispanic, Latino, or Spanish	406 (91.2)	39 (8.8)	.639
Middle Eastern or North African	76 (87.4)	11 (12.6)	.294
Native Hawaiian or other Pacific Islander	14 (100)	0 (0)	.228
White	5,086 (90.6)	526 (9.4)	.898
None of these fully describe me	75 (82.4)	16 (17.6)	**.007**
Selected more than one race and/or ethnicity	636 (89.3)	76 (10.7)	.211
Sexual orientation[a]			
Asexual	598 (89.8)	68 (10.2)	.442
Bisexual	1,692 (88.4)	221 (11.6)	**< .001**
Gay	1,908 (92.4)	157 (7.6)	**.001**
Lesbian	1,277 (90.3)	137 (9.7)	.660
Pansexual	836 (86.1)	135 (13.9)	**< .001**
Queer	2,287 (87.4)	331 (12.6)	**< .001**
Questioning	135 (82.3)	29 (17.7)	**< .001**
Same-gender loving	237 (81.2)	55 (18.8)	**< .001**
Straight/heterosexual	101 (90.2)	11 (9.8)	.874
Two-spirit	30 (78.9)	8 (21.1)	**.013**
Another sexual orientation	212 (91.8)	19 (8.2)	.537
Selected more than one sexual orientation	2,465 (87.1)	366 (12.9)	**< .001**
Income level			**< .001**
≤ $20,000	1,732 (88.3)	229 (11.7)	
$20,001 to $40,000	1,121 (90)	124 (10)	
$40,001 to $60,000	789 (91)	78 (9)	
≥ $60,001	1,500 (93.7)	101 (6.3)	
Education level			**< .001**
No high school diploma	31 (88.6)	4 (11.4)	
High school/GED graduate or some college	1,156 (87.7)	162 (12.3)	
College degree (2- or 4-year)	2,050 (90.9)	206 (9.1)	
Graduate or Professional degree	1,945 (92.4)	161 (7.6)	
Geographic region			.242
Midwest	1,117 (89.6)	130 (10.4)	
Northeast	1,150 (90.1)	126 (9.9)	
South	1,500 (90.8)	152 (9.2)	
West	1,837 (91.6)	169 (8.4)	

[a] Participants could select more than one option, thus answer choices for each identity were dichotomized to endorsed and did not endorse. For example, we compared participants who endorsed an asexual identity to those who did not endorse that specific identity. Percentage is reported as an overall total of concordance or discordance for each respective demographic characteristic.

Figure 1

Methods for assessing gender identity in The PRIDE Study's 2020 Annual Questionnaire



Method #1: Expanded Choice Model

(Multiple selection with a write-in option)

What is your current gender identity? (Check all that apply.)

□ Agender
□ Cisgender man
□ Cisgender woman
□ Genderqueer
□ Man
□ Non-binary
□ Questioning
□ Transgender man
□ Transgender woman
□ Two-spirit
□ Woman
□ Another gender identity ___________________

Method #2: Reduced Choice Model

(Single selection with reduced, predefined answer choices)

If you had to choose only one of the following terms, which best describes your current gender identity?

("Cisgender" here means identifying with the sex assigned to you at birth. For example, a cisgender woman identifies as a woman and was assigned female sex at birth.)

○ Cisgender man
○ Cisgender woman
○ Non-binary
○ Transgender man
○ Transgender woman
○ Another gender identity

Figure 2

Examples of strings associated with the five gender identity-related terms

Cisgender	Feminine	Masculine	Non-binary	Transgender
• "cis"	• "fem"	• "masc"	• "agen"	• "trans"
	• "girl"	• "boy"	• "bigen"	
	• "lady"	• "guy"	• "gray"	
	• "wom"	• "-man"	• "neu"	

References

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

Beaudry, J., Kothe, E., Singleton Thorn, F., McGuire, R., Tierney, N., & Ling, M. (n.d.). *gendercoder: Recodes Sex/Gender Descriptions into a Standard Set* [R package version 0.0.0.9000.]. Retrieved April 4, 2023, from https://github.com/ropenscilabs/gendercoder

Beischel, W. J., Schudson, Z. C., Hoskin, R. A., & van Anders, S. M. (2022). The gender/sex 3×3: Measuring and categorizing gender/sex beyond binaries. *Psychology of Sexual Orientation and Gender Diversity*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/sgd0000558

Callander, D., Newman, C. E., Holt, M., Rosenberg, S., Duncan, D. T., Pony, M., Timmins, L., Cornelisse, V., Duck-Chong, L., Wang, B., & Cook, T. (2021). The Complexities of Categorizing Gender: A Hierarchical Clustering Analysis of Data from the First Australian Trans and Gender Diverse Sexual Health Survey. *Transgender Health*, *6*(2), 74–81. https://doi.org/10.1089/trgh.2020.0050

Institute of Medicine. (2011). *The Health of Lesbian, Gay, Bisexual, and Transgender People: Building a Foundation for Better Understanding*. The National Academies Press.

Kuper, L. E., Nussbaum, R., & Mustanski, B. (2012). Exploring the diversity of gender and sexual orientation identities in an online sample of transgender individuals. *Journal of Sex Research*, *49*(2–3), 244–254. https://doi.org/10.1080/00224499.2011.596954

Lefevor, G. T., Boyd-Rogers, C. C., Sprague, B. M., & Janis, R. A. (2019). Health disparities between genderqueer, transgender, and cisgender individuals: An extension of minority stress theory. *Journal of Counseling Psychology*, *66*(4), 385.

Lombardi, E., & Banik, S. (2016). The Utility of the Two-Step Gender Measure Within Trans and Cis Populations. *Sexuality Research & Social Policy : Journal of NSRC : SR & SP*, *13*(3), 288–296. https://doi.org/10.1007/s13178-016-0220-6

Lunn, M. R., Lubensky, M., Hunt, C., Flentje, A., Capriotti, M. R., Sooksaman, C., Harnett, T., Currie, D., Neal, C., & Obedin-Maliver, J. (2019). A digital health research platform for community engagement, recruitment, and retention of sexual and gender minority adults in a national longitudinal cohort study—–The PRIDE Study. *Journal of the American Medical Informatics Association*, *26*(8–9), 737–748.

Measuring Sex, Gender Identity, and Sexual Orientation. (2022). National Academies Press. https://doi.org/10.17226/26424

Mehta, C. M., Arnett, J. J., Palmer, C. G., & Nelson, L. J. (2020). Established adulthood: A new conception of ages 30 to 45. *The American Psychologist*, *75*(4), 431–444. https://doi.org/10.1037/amp0000600

Newcomb, M. E., Hill, R., Buehler, K., Ryan, D. T., Whitton, S. W., & Mustanski, B. (2019). High Burden of Mental Health Problems, Substance Use, Violence, and Related Psychosocial Factors in Transgender, Non-Binary, and Gender Diverse Youth and Young Adults. *Archives of Sexual Behavior*. https://doi.org/10.1007/s10508-019-01533-9

Patterson, J. G., Jabson, J. M., & Bowen, D. J. (2017). Measuring Sexual and Gender Minority Populations in Health Surveillance. *LGBT Health*, *4*(2), 82–105. https://doi.org/10.1089/lgbt.2016.0026

Reisner, S. L., Biello, K., Rosenberger, J. G., Austin, S. B., Haneuse, S., Perez-Brumer, A., Novak, D. S., & Mimiaga, M. J. (2014). Using a Two-Step Method to Measure Transgender Identity in Latin America/the Caribbean, Portugal, and Spain. *Archives of Sexual Behavior*, *43*(8), 1503–1514. https://doi.org/10.1007/s10508-014-0314-2

Reisner, S. L., Conron, K. J., Scout, Baker, K., Herman, J. L., Lombardi, E., Greytak, E. A., Gill, A. M., & Matthews, A. K. (2015). "Counting" Transgender and Gender-Nonconforming Adults in

Health Research: Recommendations from the Gender Identity in US Surveillance Group. *TSQ: Transgender Studies Quarterly*, *2*(1), 34–57. https://doi.org/10.1215/23289252-2848877

Reisner, S. L., Conron, K. J., Tardiff, L. A., Jarvi, S., Gordon, A. R., & Austin, S. B. (2014). Monitoring the health of transgender and other gender minority populations: Validity of natal sex and gender identity survey items in a U.S. national cohort of young adults. *BMC Public Health*, *14*(1), 1224. https://doi.org/10.1186/1471-2458-14-1224

Ridolfo, H., Miller, K., & Maitland, A. (2012). Measuring Sexual Identity Using Survey Questionnaires: How Valid Are Our Measures? *Sexuality Research and Social Policy*, *9*(2), 113–124. https://doi.org/10.1007/s13178-011-0074-x

StataCorp. (2021). *Stata Statistical Software: Release 17. College Station, TX: StataCorp LLC.* [Computer software].

Suen, L. W., Lunn, M. R., Katuzny, K., Finn, S., Duncan, L., Sevelius, J., Flentje, A., Capriotti, M. R., Lubensky, M. E., Hunt, C., Weber, S., Bibbins-Domingo, K., & Obedin-Maliver, J. (2020). What Sexual and Gender Minority People Want Researchers to Know About Sexual Orientation and Gender Identity Questions: A Qualitative Study. *Archives of Sexual Behavior*, *49*(7), 2301–2318. https://doi.org/10.1007/s10508-020-01810-y

Tate, C. C., Ledbetter, J. N., & Youssef, C. P. (2013). A two-question method for assessing gender categories in the social and medical sciences. *Journal of Sex Research*, *50*(8), 767–776. https://doi.org/10.1080/00224499.2012.690110

The Gender Identity in U.S. Surveillance (GenIUSS) Group. (2014). *Best Practices for Asking Questions to Identify Transgender and Other Gender Minority Respondents on Population-Based Surveys* (J.L. Herman, Ed.). The Williams Institute. https://williamsinstitute.law.ucla.edu/wp-content/uploads/Survey-Measures-Trans-GenIUSS-Sep-2014.pdf

Vivienne, S., Hanckel, B., Byron, P., Robards, B., & Churchill, B. (2021). The social life of data: Strategies for categorizing fluid and multiple genders. *Journal of Gender Studies*, *0*(0), 1–16. https://doi.org/10.1080/09589236.2021.2000852

Westgard, J. O., & Groth, T. (1979). Power functions for statistical control rules. *Clinical Chemistry*, *25*(6), 863–869.