Inaccurate Estimation of Disparities Due to Mischievous Responders: Several Suggestions to Assess Conclusions

Joseph P. Robinson-Cimpian[1]

This article introduces novel sensitivity-analysis procedures for investigating and reducing the bias that mischievous responders (i.e., youths who provide extreme, and potentially untruthful, responses to multiple questions) often introduce in adolescent disparity estimates based on data from self-administered questionnaires (SAQs). Mischievous responders affect a wide range of disparity estimates, including those between adoptees and nonadoptees, sexual minorities and nonminorities, and individuals with and without disabilities. Thus, the procedures introduced here have broad relevance to research and can be widely, and easily, implemented. The sensitivity-analysis procedures are illustrated with SAQ data from youths in Grades 9–12 (*N* = 11,829) to examine between-group disparities based on sexual identity, gender identity, and physical disability. Sensitivity analyses revealed that each disparity estimated with these data was extremely sensitive to the presence of potentially mischievous responders. Patterns were similar across multiple approaches to dealing with mischievous responders, across various outcomes, and across different between-group comparisons. Mischievous responders are ubiquitous in adolescent research using SAQs and can, even in small numbers, lead to inaccurate conclusions that substantively affect research, policy, and public discourse regarding a variety of disparities. This article calls attention to this widespread problem and provides practical suggestions for assessing it, even when data are already collected.

Keywords: adolescents; disparities; equity; evaluation; lesbian, gay, bisexual, transgender, and questioning; mischievous responders; physical disabilities; questionnaires; self report; sensitivity analysis; survey research

Data from adolescents gathered through self-administered questionnaires (SAQs) are commonly used to estimate psychological and health disparities between groups such as adoptees and nonadoptees (e.g., Miller, Fan, Christensen, Grotevant, & Van Dulmen, 2000), individuals from different race/ethnicity groups (Harris, Gordon-Larsen, Chantala, & Udry, 2006), individuals with and without physical disabilities (e.g., Blum, Kelly, & Ireland, 2001; Cheng & Udry, 2002; McRee, Haydon, & Halpern, 2010), and sexual minorities and nonminorities (e.g., Bontempo & D'Augelli, 2002; Espelage, Aragon, Birkett, & Koenig, 2008; Faulkner & Cranston, 1998; Garofalo, Wolf, Kessel, Palfrey, & DuRant, 1998; LeVasseur, Kelvin, & Grosskopf, 2013; Robinson & Espelage, 2011, 2012, 2013; Russell & Joyner, 2001; Russell, Sinclair, Poteat, & Koenig, 2012). However, the use of SAQ data presents a quandary: Although SAQ methods may be best for gathering certain types of sensitive data from adolescents (particularly when anonymity is promised; Badgett, 2009; Saewyc et al., 2004; Tourangeau & Yan, 2007; Turner et al., 1998), an emerging body of literature suggests that the limitations of SAQ data can be quite severe and can lead researchers to wildly incorrect conclusions (Cornell, Klein, Konold, & Huang, 2012; Fan et al., 2006; Robinson & Espelage, 2011; Savin-Williams & Joyner, in press).[1] One prominent source of these limitations, which will also be the focus of this article, is the consistent presence of a subset of adolescent responders who provide responses they think are "funny" (e.g., reporting they are adopted when they are not, and also providing extreme responses to items on alcohol consumption, sexual activity, academics, extracurricular activities, and depression; Fan et al., 2002). This article discusses how such "mischievous responders" can lead to incorrect inferences regarding a multitude of between-group disparities (e.g., adoptee–nonadoptee, LGBQ–heterosexual), thus affecting much research on adolescents. After detailing the dangers of ignoring the presence of such

[1]University of Illinois at Urbana-Champaign

Educational Researcher, Vol. 43 No. 4, pp. 171–185
DOI: 10.3102/0013189X14534297
© 2014 AERA. http://er.aera.net

mischievous responders, a sensitivity-analysis procedure is introduced and implemented to demonstrate how researchers can assess the stability of their disparity estimates and try to reduce the influence of mischievous responders.

A few examples will illustrate the consequences and pervasiveness of mischievous responders. The first example concerns estimated disparities between adoptees and nonadoptees using data from Add Health. When completing the Wave 1 SAQ of Add Health, 458 youths reported being adopted and not living with a biological parent; on average, these youths reported substantially higher rates of alcohol and illegal-substance use (among other risky behaviors and psychological concerns) than did reported nonadoptees (Miller et al., 2000). In a subsequent paper, 88 of these 458 reported adoptee youths were found to have parents who identified their children as biological and themselves as biological parents, suggesting that 19% of youths who claimed to be adopted were falsely representing themselves as adoptees (Fan et al., 2002). When these 88 "jokesters" (in Fan et al.'s language) were removed from the dataset, nearly all of the estimated adoptee–nonadoptee disparities were substantially reduced, or even eliminated. For instance, the estimated disparity in self-esteem for true adoptees and nonadoptees was -.01 standard deviations (*SD*s), but the estimated disparity between "jokester" adoptees and true nonadoptees was -0.96 *SD*s, leading to an average estimated disparity between all (i.e., "jokester" and "true") adoptees and nonadoptees of -0.18 *SD*s. "Jokester" adoptees also provided responses to items on drinking alcohol, having physical problems, and skipping school that were 1.48, 1.67, and 1.95 *SD*s, respectively, above the means of true adoptees (Fan et al., 2002). These extreme responses by "jokester" adoptees, who made up less than 0.6% of the entire sample (but 19% of the reported adoptees), led the researchers to previously substantially overestimate the risks of adoptees (in Miller et al., 2000)—and also led to a retraction of the initial findings (Fan, 2003).

In another example, also using Add Health data, 253 youths reported on the SAQ that they used an artificial limb for a year or more, but in-person follow-up interviews revealed that only two of these youths used an artificial limb—a full 99% provided inaccurate responses to the artificial-limb question on the SAQ (Fan et al., 2006), which has implications for research findings based on these data (e.g., Blum et al., 2001; Cheng & Udry, 2002; McRee et al., 2010). In addition, youths who misrepresented their adoptee or disability status were also much more likely to misrepresent their foreign-born status, gender, age, and race/ethnicity (Fan et al., 2006), which has implications for studies examining disparities along these dimensions as well. Although some misreporting may be due to error or confusion, the consistent patterns of misreporting just described, combined with patterns of extreme reporting of risky behavior, suggest that these youths *intentionally* misreported their data (Fan et al., 2006), hence the term *mischievous responders*. Moreover, mischievous responders were disproportionately concentrated among individuals reporting minority-group status (e.g., adoptees, individuals with disabilities), thereby maximizing the influence such mischievous responders have on disparity estimates due to their relatively higher proportions among reported minority-group members than among reported majority-group members. Quite simply, mischievous responders can lead to biased estimates on a wide range of between-group disparities, particularly when studying underrepresented minorities.

Although it is perhaps easier to identify inaccurate responses in the Add Health data because of the various methods of data collection used by this survey (which included in-person interviews and interviews with parents), Add Health's is by no means the only SAQ on which youths provide inaccurate responses. [In fact, it is to Add Health's credit that it has multiple methods of measurement (i.e., student SAQs, parent SAQs, in-home visits, and school records) that allow for this kind of triangulation that Fan et al. (2002, 2006) capitalized on to ingeniously identify mischievous responders.] Indeed, other research found that simply asking youths at the end of a SAQ whether they answered the questions truthfully resulted in up to 12% reporting that they did not (Cornell et al., 2012). But what of other studies that used SAQs and did not have a way of assessing truthfulness? Are we to assume that youths are completely honest on those SAQs? In reality, that is what researchers implicitly do anytime they do not assess the sensitivity of their estimates to the presence of potentially mischievous responders.

The solutions to this problem that have been proposed so far are, under most circumstances, impractical. For example, using other sources of data to confirm SAQ-reported data (e.g., using data from in-person observations or parents; Fan et al., 2006) is often not feasible and even unethical for some disparities of interest (e.g., asking parents of LGBQ teens to confirm that their children are in fact LGBQ may put the children at risk; D'Augelli & Grossman, 2006). Or, to take another example, adding items to the SAQ that ask respondents if they provided truthful responses to earlier items involves adding new items to a SAQ (Cornell et al., 2012; see also Poulin, MacNeil, & Mitic, 1993, who added a fictitious drug to a list of real illegal drugs), which is not possible if researchers have already collected their data. In contrast to these prior approaches, this article will focus on sensitivity analyses that researchers can implement with *items already existing* in their SAQ data, which therefore may be much broader in their applicability than other methods suggested in the literature.

Methods

This article introduces a four-step sensitivity-analysis procedure that researchers can implement to assess the validity of their estimates (see Figure 1). In brief, the strategy of this four-step procedure is to identify youths who systematically provide unusually high numbers of low-frequency responses (e.g., reporting they are blind *and* deaf *and* in a gang *and* parenting multiple children), and then to compare the estimated disparities when including and excluding these multiple low-frequency responders. Step 1 requires researchers to identify a set of SAQ items that permitted youths to provide a response that adolescents may have considered "funny" (e.g., claiming they are blind; reporting they have two or more children of their own; stating they are in a gang) but that are in principle unrelated (or inversely related) to group identification/status. This collection of items is termed the *screener*. In Step 2, screener index values are calculated using a screener-indexing approach (discussed below). Step 3 involves examining the representation of groups throughout different ranges of index values (e.g., Are youths reporting minority-group

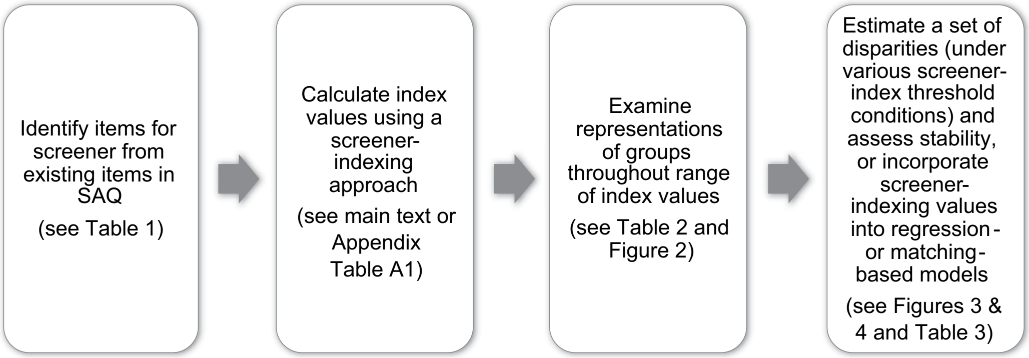
Identify items for screener from existing items in SAQ (see Table 1)
Calculate index values using a screener-indexing approach (see main text or Appendix Table A1)
Examine representations of groups throughout range of index values (see Table 2 and Figure 2)
Estimate a set of disparities (under various screener-index threshold conditions) and assess stability, or incorporate screener-indexing values into regression- or matching-based models (see Figures 3 & 4 and Table 3)

FIGURE 1. *Sensitivity-analysis procedure*
This procedure can be applied with any of screener-indexing approaches described in the main text or Appendix Table A1.

status overrepresented in a certain range of index values, such as the most extreme index values?). Disparities are estimated and their stability is assessed in Step 4. More precisely, a *set* of disparities is estimated with various methods of screening out respondents. For example, a researcher interested in a minority–majority disparity in sense of school belongingness would first estimate the disparity among all respondents, and then estimate the same disparity if youths providing extreme responses on the screener items were removed. Then, researchers assess the stability of their disparity estimates under the various conditions of screening out respondents. This stability can then be factored into their interpretation of their results (e.g., the more unstable the disparities, the stronger the possibility of bias in estimates). An alternative, regression-, or matching-based approach is also discussed.

Step 1 (Identify Screener Items)

When choosing screener items in Step 1, it is important to select SAQ items that are in principle unrelated or inversely related to group identification/status (but may be positively correlated in the SAQ data, likely due to the mischievous responders). For example, a SAQ may ask teenage respondents how many children they have, with possible responses of *none, 1*, or *2 or more*. Youths may think it is "funny" to select the option of *2 or more* children. And one would not expect reported lesbian/gay-identified individuals to respond that they have *2 or more* children more often than reported heterosexual-identified adolescents, but this is precisely what the SAQ data suggest (12.1% vs. 1.3%; or 6.6% for LGBQ-identified as a whole vs. 1.3% for heterosexual-identified, as shown in Table 1). Thus, this number-of-children item may be a good candidate for a screener item if one were examining lesbian/gay–heterosexual disparities, as I will do in an empirical illustration below. (Table 1 provides a list of all screener items used for the empirical analysis below and the percentage of each group providing each low-frequency response option.) However, if one is exploring disparities between, say, teen parents and nonparents, then the question about number of children could not be included in the screener. Hence, the selection of items for inclusion in the screener will be specific to the groups examined and will be a function of the items that are available. To further illustrate this point, one would not expect transgender youths to be blind more often than cisgender youths (see Table 1, 31.4% vs. 3.1%), so an item asking youths if they are blind may be a good screener item for transgender–cisgender disparities. Conversely, the blindness item is an *unsuitable* screener item for disability-based disparities, and thus it will not be used in the screener when I compare youths with and without physical disabilities (this is reflected in Table 1 as well).

Step 2 (Calculate Screener-Indexing Values)

I introduce two different approaches for calculating the screener-indexing values (see Appendix A for additional details). These approaches differ in how they aggregate respondents' responses on the screener items: The first approach I discuss is a *count-based* screener-indexing approach, which counts the number of low-frequency responses (Robinson & Espealge, 2011). For example, if there are 10 screener items, then youths will have a discrete value from 0 (if they provide no low-frequency responses) to 10 (if they only provide low-frequency responses) on this measure.

The second measure is a *probability-based* measure of the prevalence of low-frequency responses on screener items. In this approach, individual *i*'s value of *P* is the product of *i*'s response probabilities *p* for each item *m* in a group of *M* items: $P_i = \sum_{m=1}^{M}(p_{im})$. For a simple illustration, if there were only two screener items (i.e., M = 2), and one individual gave a response that 10% of individuals provided for Item 1, and a response for Item 2 that 2% of individuals provided, his value of P_i = .002; whereas an individual who provides responses for Items 1 and 2 that 90% and 98% of respondents provided would have P_i = .882. The values of *P* are then ranked. The advantage of this second approach is that it allows one to weight the lower low-frequency responses more than the higher low-frequency responses, whereas the first (i.e., count-based) approach weights all low-frequency responses equally. For example, we can see from Table 1 that reporting having two or more children has a lower probability than reporting having a family member in a gang; thus, this second screener-indexing approach allows researchers to give additional weight to reporting more rare events. Additional screener-indexing approaches are possible, and were in fact developed for this study. These additional approaches yielded similar results to the approaches just mentioned, and thus for brevity they are not discussed here, but the details of these approaches can be found in Appendix A.

Table 1
Screener Items: Survey Items Used to Identify Multiple Low-Frequency Responders and Percentage of Each Reported Group Providing Each Low-Frequency Response

	Sexual Orientation		Gender Identity		Physical Disability	
Item (*and Low-Frequency Response*)	Heterosexual (*N* = 11,058)	LGBQ (*N* = 771)	Cisgender (*N* = 11,625)	Transgender (*N* = 204)	Not Disabled (*N* = 11,528)	Disabled (*N* = 301)
1. Provided a height in the top or bottom 2.5%	3.7%	13.5%	3.7%	41.7%	N/U (3.6%)	N/U (31.9%)
2. Provided a weight in the top or bottom 2.5%	4.0%	13.1%	4.1%	30.4%	N/U (3.9%)	N/U (30.6%)
3. Are you deaf or have a hearing impairment? (*Yes*)	1.3%	8.9%	1.3%	25.5%	N/U (0.9%)	N/U (35.2%)
4. Are you blind or have vision impairment? (*Yes*)	2.9%	13.9%	3.1%	31.4%	N/U (2.7%)	N/U (39.5%)
5. When was the last time you visited a dentist? (*3 or more years ago*)	3.4%	12.8%	3.5%	30.9%	3.4%	25.2%
6. How many times have you been pregnant or have gotten a girl pregnant? (*2 or more times*)	0.7%	7.8%	0.7%	25.0%	0.7%	18.3%
7. How many children do you have? (*2 or more*)	1.3%	6.6%	1.3%	23.0%	1.2%	17.6%
8. Is one or more of your family members in a gang? (*Yes*)	3.1%	15.2%	3.2%	41.7%	3.2%	28.9%
9. Are you in a gang? (*Yes, currently*)	1.9%	11.3%	1.9%	35.3%	1.9%	27.6%
10. In the past month, how many days have you carried a weapon to school? (*6 or more days*)	1.1%	9.2%	1.2%	28.4%	1.1%	22.3%

Note. The total number of screener items for the LGBQ–heterosexual and transgender–cisgender analyses is 10. The total number of screener items for the disabled–nondisabled analyses is six because the items on height, weight, blindness, and deafness were not used due to possible valid correlations with reported disability. "N/U" in the physical disability columns refers to "not used," as in "not used in the screener"; however, the values are presented here so that readers can see them anyway. Finally, the weighted percentages in the first row do not sum to the expected 5% because observations with values at the 2.5th and 97.5th percentiles for height were not flagged as low-frequency responses. The same is true for the second row (weight). Supplemental analyses were conducted to test how including observations at the 2.5th and 97.5th percentiles affects the patterns reported in the article. These analyses revealed nearly identical patterns. Thus, flagging these borderline observations for height and weight did not lead to different patterns or conclusions.

Steps 3 and 4 of the sensitivity-analysis procedure will be elaborated on in the context of an empirical example, introduced next.

Data

Although the sensitivity-analysis and screener-indexing approaches are *general* and thus can be used in a variety of contexts, to illustrate the approaches I will use data from the 2012 Dane County Youth Assessment (DCYA; for examples of research using the 2008–2009 DCYA, see Poteat, Mereish, DiGiovanni, & Koenig, 2011; Robinson & Espelage, 2011, 2012, 2013; Russell et al., 2012) to assess the sensitivity of estimated (1) LGBQ–heterosexual disparities, (2) transgender–cisgender disparities, and (3) physically disabled–nondisabled disparities for three outcomes: (1) frequent suicidal ideation (i.e., "almost all the time") in the past month, (2) school belongingness (a composite of six items; α = .81), and (3) cocaine/crack use in the past year. (See Appendix B available on the journal website for more details on each item and scale.)

The DCYA is an anonymous Web-based survey administered in school in January and February of 2012 to students in high school in Dane County, Wisconsin. The final analytic dataset contains a total of 11,829 students in 22 schools. The survey assessed a wide range of psychological and health indicators, as well as various attitudes and social behaviors. Students completed these anonymous surveys independently while in school during proctored sessions. A waiver of active consent was employed, and child written assent was used. University institutional review board (IRB) approval was obtained. Surveys were given to all students in non-Madison schools and were given to (randomly sampled) half of students in Madison schools. The response rate was very high: in each of the 22 schools, over 90% of students surveyed provided responses. Additional details on the data and exact questionnaire-item phrasing and coding can be found in Appendix B available on the journal website.

In the demographic section of the survey, students were asked to identify their sexual orientation as straight/heterosexual, gay or lesbian, bisexual, or questioning their sexual orientation. Students could choose only one category. Shown in Tables 1 and 3, 771 students (6.5%) selected a LGBQ category. Students were also asked whether or not they identified as transgender, and 204 (1.7%) students chose yes. In a later portion of the survey, students were asked whether they had a physical disability, and 301 (2.5%) reported they did.

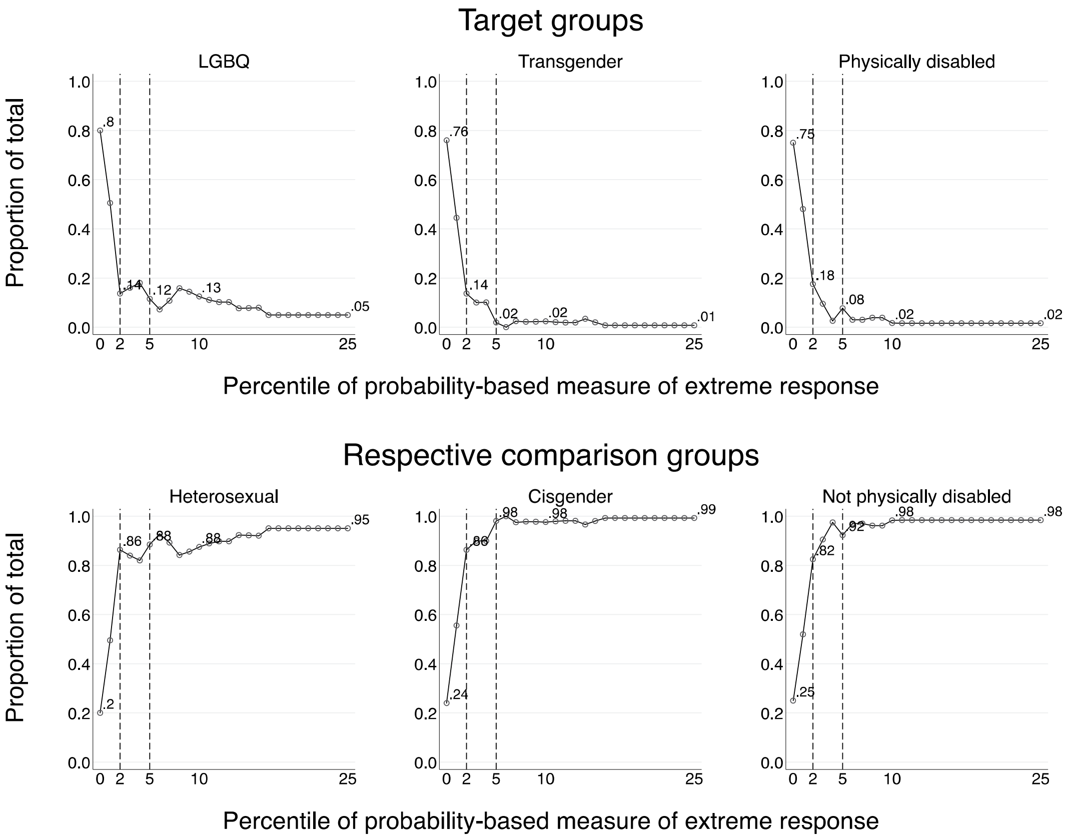

FIGURE 2. *Proportion of youths reporting a given group, by percentile range*

Each panel in the figure presents the proportion of youths who reported a given group affiliation within a specified percentile range. The ranges presented are the top 0.2 percentile (i.e., respondents with the very most extreme response patterns on the screener), followed by the range from percentiles 0.2 to 1, then percentiles 1 to 2, percentiles 2 to 3, and so on, up to percentiles 24 to 25. For example, the upper left panel illustrates that 80% of respondents in the top 0.2 percentile reported LGBQ identities, while 14% did so between the percentiles 1 and 2, and 13% did so between the percentiles 9 and 10.

Disparities were estimated using linear models or linear probability models, for ease of interpretation and comparability of scale across models with differing samples (Mood, 2010), with cluster-robust standard errors to account for the nesting of students within schools. To further facilitate interpretation (and to make results more comparable with prior studies, discussed later in the Discussion section), results are reported in standard deviation units and can be interpreted as effect sizes (i.e., Cohen's *d*; Cohen, 1988).

Results

Step 3 (Examine Representations of Groups at Various Index Values)

If we find that a group is represented in roughly the same proportion at all screener-index values, then we might not suspect mischievous responders to be disproportionately concentrated among that group; however, if we see that the representations of groups change across the index values, then we may become more suspicious. In the DCYA data, youths who reported LGBQ identification on the SAQ were disproportionately represented among respondents providing multiple low-frequency responses, as were youths who reported they identified as transgender and youths who reported having a physical disability.

In Table 2, over 95% of respondents provided less than two low-frequency responses, and about 2% of respondents provided three or more low-frequency responses. However, disaggregating by reported group identification, we see that although only 1.5% of reported heterosexual-identified youths provided three or more low-frequency responses, over 11.7% of reported LGBQ-identified youths provided three or more low-frequency responses. The discrepancies by reported gender identity and disability are even more striking. For instance, 1.5% of reported cisgender-identified youth provided three or more low-frequency responses, whereas nearly 40% of reported transgender-identified youth provided at least three low-frequency responses. Figure 2 shows similar information, but from the probability-based screener-indexing approach. For example, 80%, 76%, and 75% of youths who provided the most extreme response-patterns (i.e., the top 0.2%) to the screener items also reported being LGBQ, transgender, and physically disabled, respectively. Among less extreme response-patterns (i.e., looking at the right edge of each panel in Figure 2), we see that about 5%, 1%, and 2% of youths reported being LGBQ, transgender, and disabled—estimates that are more in line with population estimates. These patterns

Table 2
Cell Sizes and Percentages, by Reported Group Affiliation and Count of Low-Frequency Responses Provided

	Heterosexual		LGBQ		Cisgender		Transgender		Not Physically Disabled		Physically Disabled	
	N	***Col. %***	***N***	***Col. %***	***N***	***Col. %***	***N***	***Col. %***	***N***	***Col. %***	***N***	***Col. %***
Total	11,058	100.00	771	100.00	11,625	100.00	204	100.00	11,528	100.00	301	100.00
Number of low-frequency responses												
0	9,342	84.48	484	62.78	9,755	83.91	71	34.80	10,584	91.81	176	58.47
1	1,283	11.60	153	19.84	1,408	12.11	28	13.73	725	6.29	33	10.96
2	268	2.42	44	5.71	287	2.47	25	12.25	133	1.15	13	4.32
3	70	0.63	10	1.30	73	0.63	7	3.43	43	0.37	19	6.31
4	31	0.28	7	0.91	31	0.27	7	3.43	17	0.15	24	7.97
5	18	0.16	10	1.30	18	0.15	10	4.90	13	0.11	7	2.33
6	12	0.11	10	1.30	13	0.11	9	4.41	13	0.11	29	9.63
7	17	0.15	11	1.43	19	0.16	9	4.41				
8	10	0.09	16	2.08	12	0.10	14	6.86				
9	3	0.03	8	1.04	4	0.03	7	3.43				
10	4	0.04	18	2.33	5	0.04	17	8.33				

Note. The total number of screener items for the LGBQ–heterosexual and transgender–cisgender analyses is 10. The total number of screener items for the disabled–nondisabled analyses is six because the items on height, weight, blindness, and deafness were not used due to possible valid correlations with reported disability.

shown in Table 2 and Figure 2 are consistent with more potentially mischievous responders among youths reporting LGBQ identities, transgender identities, and physical disabilities. Importantly, this does not imply that *actual* LGBQ- or transgender-identified youths or youths having physical disabilities are mischievous, but rather that youths *reporting* sexual minority identities or physical disabilities—regardless of actual identity or disability status—are potentially more mischievous, just as was seen with true and "jokester" adoptees in earlier studies (cf., Fan et al., 2002, 2006).

Step 4 (Estimate Disparities and Assess Stability)

I begin by presenting the transgender–cisgender disparity in suicidal ideation, first for the count-based screener-indexing method, then for the probability-based screener-indexing method; then I will present other between-group disparities. Among the full sample (i.e., before screening out anyone), I estimate that nearly 2 *SD*s more (d = 1.98, SE = 0.32, p < .001; see top middle panel of Figure 3) of transgender-identified youths (25.4%) often thought about killing themselves compared to cisgender-identified youths (1.2%). Screening out only those who provided three or more low-frequency responses (i.e., less than the top 2%), the estimated transgender–cisgender disparity is eliminated (d = 0.00, SE = 0.06, p = .98), with 0.8% of each group reporting thinking often about suicide. Similar patterns are seen in the continuous probability-based measure (Figure 4). Screening out individuals whose responses were in the top 2% of extreme-response patterns leads to a reduction in the estimated transgender–cisgender suicidal ideation disparity from 1.98 *SD*s (SE = 0.32, p < .001) to 0.00 *SD*s (SE = 0.06, p = 1.00).

Turning to the other between-group disparities in frequent suicidal ideation (columns 1 and 3 of Figures 3 and 4), we again see that using a screener created from existing SAQ items results in reductions in the estimated disparities. For example, using the probability-based approach (Figure 4), removing the top 2% of extreme responders reduces the estimated LGBQ–heterosexual disparity by more than two thirds, from d = 0.74 (SE = 0.11, p < .001) to d = 0.22 (SE = 0.05, p < .001), and reduces the estimated disabled–nondisabled disparity by over 80%, from d = 1.88 (SE = 0.23, p < .001) to d = 0.35 (SE = 0.13, p < .001). Similar patterns can be seen for school belongingness and cocaine/crack use in the past year.

If the screener is valid (i.e., the items are in principle unrelated or inversely related with group identification/status), then such sudden drops call into question the validity of the results based upon the full sample. Thus, for instance, the LGBQ–heterosexual estimated disparity of 12 percentage points (d = 0.75, SE = 0.11, p < .001) in cocaine/crack use in the past year is likely an overestimate driven by mischievous responders, and a more plausible estimate is 4 percentage points (d = 0.25, SE = 0.06, p < .001)—an estimate associated with removing the top 2% of extreme responders. If one bears in mind that prior studies suggest up to 12% of adolescents provide untruthful responses (Cornell et al., 2012), removing 2% does not seem overly conservative. Note, however, that although the disparities examined here reduced in magnitude, many do not reduce to nonsignificant levels and their standardized differences remain sizable. For instance, the persistence of the moderately sized transgender–cisgender disparity in cocaine/crack use (d = 0.38, SE = 0.17, p = .04) when even the top 25% of extreme responders are removed—an extremely conservative approach to screening out potentially mischievous responders—suggests that this disparity is real and that perhaps action should be taken to address this disparity.

The Step 4 approach above requires researchers to provide a *set* of estimates for the various screening thresholds, as presented in Figures 3 and 4. Some researchers may find it overwhelming to estimate and provide a large set of estimates for each disparity

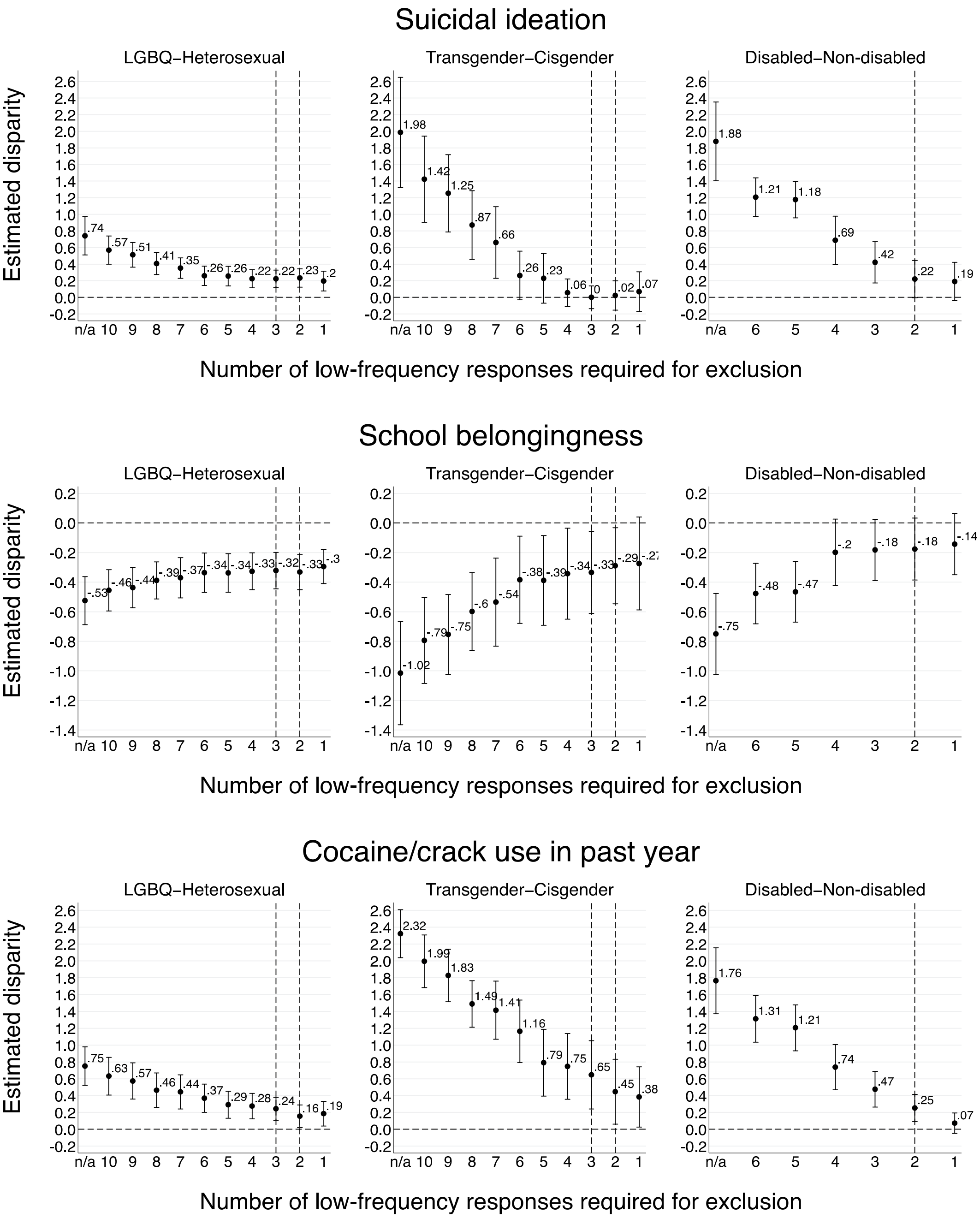

FIGURE 3. *Estimated disparities, by outcome, group comparison, and count-based exclusion criteria*
Disparities are in *SD*s. Cluster-robust 95% confidence intervals are presented as bars around the point estimates. If the 95% confidence interval contains 0, the disparity is not statistically significant ($p > .05$). "n/a" refers to the full sample being used to estimate the disparities. For ease of comparison with Figure 4, the first vertical dashed line in the first two columns corresponds to the top 2% of observations being removed, and the second dashed line refers to the top 5% removed; the sole vertical line in the final column refers to 5% removed.

and thus may prefer an approach that only requires a *single* estimate per disparity. Rather than estimate a set of values and assess stability, researchers can simply use the screener-indexing values as *covariates* (using a regression model) or as variables to *match* on (using a matching-based estimator). Intuitively, when researchers do *not* adjust for mischievousness (or proxies for it, such as the screener-index values), then disparity estimates may suffer from omitted variable bias. In this case, the omitted variable is mischievousness, which is confounded with both group identification and the outcome of interest. By including variables that proxy for mischievousness in a regression model along with the group identification variable(s), researchers can begin to address concerns of omitted variable bias. This regression-based alternative to the set-based approach is implemented in Table 3. In general, the regression-based analyses yielded estimates that are similar to estimates obtained by screening out the top 1% to 5% of extreme responders. For example, the LGBQ–heterosexual disparity in frequent suicidal ideation reduces from 9 percentage

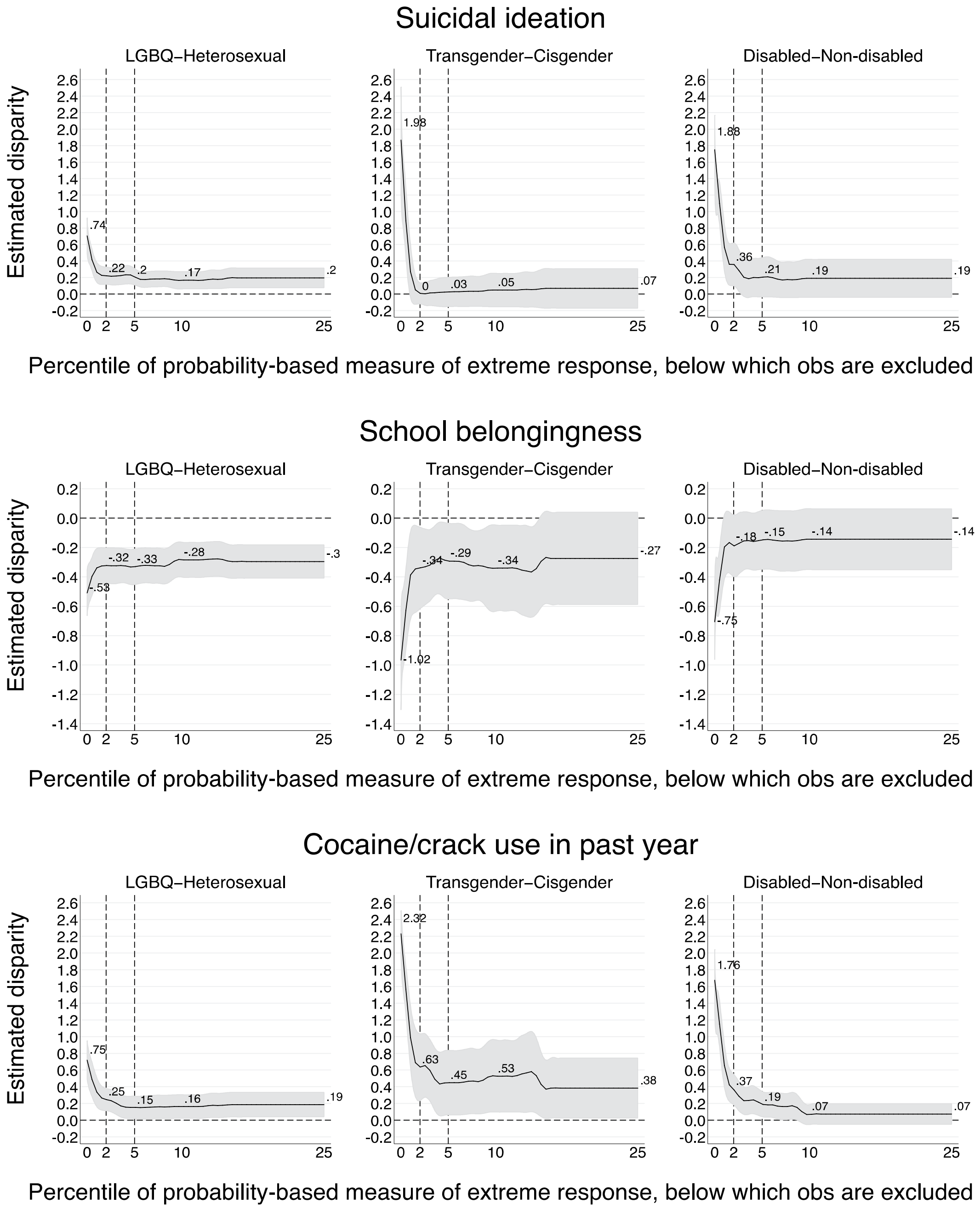

FIGURE 4. *Estimated disparities, by outcome, group comparison, and probability-based exclusion criteria*
Disparities are in *SD*s. Cluster-robust 95% confidence intervals are presented as bands around the point estimates. If the 95% confidence interval contains 0, the disparity is not statistically significant ($p > .05$).

points (d = 0.74, SE = 0.11, $p < .001$) to 2 percentage points (d = 0.23, SE = 0.05, $p < .001$) when statistically adjusting for a count-based measure in a regression framework.

Discussion

This article highlights an important aspect of research on disparities that is often overlooked—namely, the bias that is introduced by mischievous responders. After describing previous examples of biasing effects for a range of between-group disparities, I introduced new ways to assess the sensitivity of disparity estimates to the presence of potentially mischievous responders. These sensitivity analyses can be implemented with existing data (rather than requiring additional data collection), which makes them feasible in a wide range of circumstances. Across different screener-indexing approaches, common patterns emerged: Removing a very small subset of the sample (just the most extreme 1 to 2% on the screener) resulted in substantially different estimated disparities. An alternative approach was also introduced to account for mischievous responders in a regression-based or matching-based analysis; the results from the alternative approach were consistent with estimates associated with removing the top 1 to 5% on the screener. Moreover, similar patterns of reductions were seen in three different between-group

Table 3
Regression-Based Estimated Disparities, by Outcome, Group, and Screener-Indexing Approach Used as Covariates

	No Additional Covariates	Count-Based as Covariate	Probability-Based as Covariate
Suicidal ideation			
LGBQ–heterosexual	0.740 (0.111), $p < 0.001$	0.232 (0.050), $p < 0.001$	0.236 (0.051), $p < 0.001$
Transgender–cisgender	1.985 (0.318), $p < 0.001$	0.150 (0.138), $p = 0.289$	0.153 (0.158), $p = 0.342$
Disabled–nondisabled	1.877 (0.228), $p < 0.001$	0.641 (0.120), $p < 0.001$	0.706 (0.111), $p < 0.001$
School belongingness			
LGBQ–heterosexual	-0.525 (0.078), $p < 0.001$	-0.294 (0.056), $p < 0.001$	-0.294 (0.057), $p < 0.001$
Transgender–cisgender	-1.015 (0.168), $p < 0.001$	-0.225 (0.141), $p = 0.125$	-0.225 (0.146), $p = 0.138$
Disabled–nondisabled	-0.750 (0.132), $p < 0.001$	-0.176 (0.089), $p = 0.060$	-0.210 (0.087), $p = 0.025$
Cocaine/crack use			
LGBQ–heterosexual	0.750 (0.110), $p < 0.001$	0.242 (0.064), $p = 0.001$	0.252 (0.067), $p = 0.001$
Transgender–cisgender	2.322 (0.137), $p < 0.001$	0.491 (0.147), $p = 0.003$	0.518 (0.148), $p = 0.002$
Disabled–nondisabled	1.762 (0.188), $p < 0.001$	0.508 (0.106), $p < 0.001$	0.512 (0.114), $p < 0.001$

Note. Cluster-robust standard errors appear in parentheses. Each regression for column “No additional covariates” contains only the LGBQ, transgender, or disability indicator variable as a covariate. Each regression for “Count-based as covariate” contains the group indicator and an indicator for each number of low-frequency responses (i.e., an indicator for one low-frequency response, an indicator for two low-frequency responses, and so on). Each regression for “Probability-based as covariate” contains the group indicator and a fourth-order polynomial for the continuous function of P (i.e., linear, quadratic, cubic, and quartic terms for the natural log of P).

comparisons: those examining LGBQ–heterosexual disparities, those examining transgender–cisgender disparities, and those examining disabled–nondisabled disparities. These results demonstrate that (1) a very small group of potentially mischievous responders can dramatically alter our impressions of disparities and (2) mischievous responders can affect numerous types of between-group comparisons. The results also imply that the more important analytic choice is *whether or not* to conduct a sensitivity analysis rather than which screener-indexing methods to use for this analysis, as many of these methods behaved similarly.

Potential Extensions of This Work

This article provides examples of how potentially mischievous responders can affect disparities along the dimensions of sexual identity, gender identity, and disability status; however, mischievous responders are not confined to these dimensions, extending to numerous others such as race/ethnicity. For example, the DCYA data suggest that youths who reported Middle Eastern racial/ethnic identities are more than 700 times as likely as reported non–Middle Eastern White youths to be blind *and* deaf *and* in a gang *and* have more than two children *and* have an extreme height and weight *and* not see a dentist, etc., *all at the same time.* (It may come as no surprise to learn that these youths also reported they were LGBQ-identified *and* transgender *and* physically disabled too.) As has been argued in this article, these youths are likely to be mischievous responders who think it is funny to claim a Middle Eastern identity. The problem for education researchers is that the presence of such responders throws off estimates of the school belongingness and engagement of youths with Middle Eastern identities.

The focus of this article is on how mischievous responders can lead to inaccurate estimates of disparities, but their biasing effects extend to other estimates, such as mediation-based estimates. For example, researchers may theorize that LGBQ–heterosexual disparities in substance use are partially mediated by LGBQ youths’ higher prevalence of reported sexual abuse (see, e.g., Marshal et al. 2008; Saewyc et al., 2006). If, however, mischievous responders think it is “funny” to say they are sexually abused, then the mediator is conflated with mischievousness as well (Baron & Kenny, 1986; Imai, Keele, Tingley, & Yamamoto, 2011). In this case, researchers are left with both a biased *direct* pathway (from LGBQ to substance use, not via sexual abuse) and *indirect* pathway (from LGBQ to sexual abuse to substance use). Thus, mischievous responders may not only lead us to incorrect inferences about disparities, but they may also lead us to incorrect inferences about the *mechanisms* by which these disparities may arise.

Methodological extensions that build on the premise of the sensitivity analyses introduced here are also possible. One possible extension is with respect to different screener-indexing approaches. This article discusses two basic indexing approaches—a count-based approach and a probability-based one. These approaches are intuitive and easy to implement; however, more complex approaches (e.g., involving structural equation modeling) can also be used to identify mischievous responders. For example, researchers can use structural equation modeling to predict the observed responses to the screener items as a function of the latent construct of mischievousness (here, assuming mischievousness is a *continuous* measure). Using the resulting model estimates, researchers can predict mischievousness factor scores for each individual, to be ranked and used as the screener index in Step 2. Another possible extension is to use latent class analysis to predict from the observed responses to the screener items which of two latent *discrete* classes the individual is a member of: mischievous responders or nonmischievous responders. Individuals predicted to be mischievous could then be screened out and the estimate stability assessed. These more complex models may yield greater sensitivity in detecting mischievous responders; however, more complex models may be less transparent. Future research should explore circumstances under which greater model complexity improves the ability to identify mischievous responders, which may justify the tradeoff with transparency.[2]

Strengths, Limitations, and Other Considerations

As previously noted, the approaches proposed in the current article have two major advantages over previously suggested methods. First, the presently proposed approaches do not require additional data collection, which may make these approaches more practical, cost-effective, and applicable to already collected data. Second, the current approaches may be feasible in situations where previously suggested methods are not, for a variety of reasons: For example, it is unethical to ask parents about their children's reported sexual identity, which makes the triangulation method (Fan et al., 2006) infeasible.

Nevertheless, the methods proposed in this article have several limitations. The principal limitation of this approach is that one cannot be certain that individuals who provided extreme responses were in fact being untruthful. A related (but opposite) concern involves creating a screener that fails to identify mischievous responders. Thus, the screener must be sensitive enough to measure mischievousness reasonably well, neither over- nor underidentifying it. To partially address this limitation, researchers can create different sets of items to constitute the screener; similar patterns of estimate (in)stability across the different screeners may increase confidence in the sensitivity analysis.[3] To increase the likelihood that the screener is correctly identifying mischievous youths, researchers should use prior research and theory to construct a screener comprised of items that are in principle unrelated or inversely related to group identification. Thus, although one cannot be certain of which respondents are mischievous, confidence can be increased by sound choices in screener-item selection. Nevertheless, it is important to bear in mind that screening out observations believed to be mischievous responders rests on assumptions about the ability of the screener to correctly identify such responders, as discussed above. Thus, if researchers choose to screen out observations after conducting sensitivity analyses, this analytic choice—as well as the assumption about the validity of the screener—should be clearly stated. Finally, the screener cannot circumvent the fact that the SAQ data are self-reported. For instance, some individuals who truly identify as LGBQ may be reluctant to report this on a questionnaire (Coffman, Coffman, & Ericson, 2013), and the screener cannot uncover which individuals these are. Likewise, youths may misreport information on items of interest (e.g., drug use; Poulin et al., 1993; Tourangeau & Yan, 2007), and thus disparity estimates will only be able to reflect what is reported and not necessarily what is true. However, sensitivity analyses may reduce the likelihood that deliberately false responses are included in estimates.

Despite these limitations, researchers must consider the consequences of the alternative, which is typically *not* conducting sensitivity analyses when working with anonymous SAQ data. Failure to conduct such sensitivity analyses may lead to false confidence in inaccurate estimates of between-group disparities, which—beyond generating inaccurate research—can lead to ineffective policymaking and is also likely to perpetuate negative stereotypes about marginalized groups. For example, a recent meta-analysis found an average effect size of 0.72 for LGBQ–heterosexual disparities in recent cocaine/crack use among youths (Marshal et al., 2008). This estimate is remarkably close to (and statistically indistinguishable from) the estimate of d = 0.75 found in the current study when using the full sample; however, it is meaningfully (and statistically) different from the estimate of d = 0.25 obtained by removing just the top 2% of extreme responders. This result is noteworthy not only from a scientific standpoint but also from a broader societal perspective: Sensitivity analyses of the sort suggested here may have prevented well-intentioned research from being used by groups who claim that LGBQ identification leads to substance abuse and other risky behaviors and who thus advocate *against* using school resources to help LGBQ youths (Anderson, n.d., 2008; Glenn, 2002). Furthermore, inflated estimates of disparities may perpetuate stigma against minority groups (Hatzenbuehler, 2009; Meyer, 2003), and may focus undue attention on (inaccurate) negative aspects of minority-group identification/status, rather than on the resiliency demonstrated by members of minority groups (Russell, 2005; Saewyc, 2011; Savin-Williams, 2001). Thus, there is great potential for both scientific and practical damage by not performing sensitivity analyses to assess data validity and ensure robust results.

Within the past decade, much research has been criticized for failure to replicate and for exaggerated results (for critiques, see, e.g., Fanelli & Ioannidis, 2013; Gelman, in press; Ioannidis, 2005a, 2005b). The methods proposed in the current article may serve as another safeguard against the generation of inaccurate research findings. Importantly, note that I do *not* mean to imply that researchers using SAQ data have *intentionally* exaggerated the magnitude of their findings. Quite the contrary, most researchers probably do not suspect that their data are susceptible to such dangers. Thus, the goals of this article are to raise awareness of the pervasiveness of mischievous responders and to provide solutions for assessing the possibility of bias due to such responders.

Finally, this research relates to the broader topic of cross-methodology research (for a discussion, see Moss et al., 2009). Sensitivity analyses such as those proposed in the current article may serve to reduce the influence of mischievous responders and allow quantitative findings to be more easily related to qualitative findings, where mischievousness might be less of a concern (cf., Fan et al., 2006). Although findings from qualitative and quantitative studies could diverge for various reasons (e.g., different selection criteria, error due to random sampling or small samples), all else equal they should converge. Sensitivity analyses may facilitate this convergence by helping to identify mischievous responders who might disproportionately distort estimates on minority populations from large-scale survey data. For example, estimates from nonscreened SAQ data suggest a dire situation for LGBQ youths, with little to no suggestion that LGBQ–heterosexual gaps have diminished historically (Saewyc et al., 2008); however, as was demonstrated in the current article, such gap estimates are likely sensitive to the presence of mischievous responders, and removing the most likely mischievous responders substantially reduced the magnitudes of these estimates. Interestingly, the resulting smaller disparity estimates are more compatible with some in-depth, *qualitative* accounts that suggest that sexual-minority identification is becoming less of a stigmatizing factor among adolescents than it was in past years (McCormack, 2011; Savin-Williams, 2005). In this case, the sensitivity analyses revealed that the SAQ-based estimates may actually be more in line with recent qualitative evidence than previously thought. If the various forms of evidence suggest similar disparities, then perhaps we can use the qualitative evidence to unpack the experiences behind the large-survey-based estimates, working beyond methodological boundaries to ultimately gain deeper insights into the lives of adolescents.

Conclusion

The presence of a small group of mischievous responders can have a dramatic effect on disparity estimates, as was demonstrated here with disparities based on sexual identity, gender identity, and physical disability, as well as in other empirical studies on adoption, physical disability, and foreign-born status (Fan et al., 2002, 2006). The sensitivity-analysis methods introduced in this article may serve as easily implementable checks on the validity of the conclusions drawn concerning a broad range of adolescent disparities. The consistent application of such sensitivity analyses is likely to improve our ability to produce sound research that enhances effective policymaking for adolescent well-being.

NOTES

This research was partially funded through a fellowship from the National Academy of Education/Spencer Foundation. I thank Andrei Cimpian for helpful comments throughout the writing of this article. I also thank Dan Barry, Dorothy Espelage, Sarah Lubienski, and Sean Reardon for helpful comments on previous drafts. I thank K-12 Associates and the Dane County Youth Commission (Dane County, Wisconsin) for sharing their data.

[1] The references in the main text focus primarily on adolescents who are suspected of providing mischievous responses. However, there is a broader literature on false reporting on sensitive items in questionnaires. For further reading, see Glynn (2013), Kreuter, Presser, and Tourangeau (2008), Tourangeau and Yan (2007), and Wolter and Preisendörfer (2013).

[2] Two supplemental analyses were performed to preliminarily assess the need for more complex models. The first supplemental analysis used structural equation modeling as a screener-indexing approach. With the DCYA data, the structural-equation-modeling-based indexing approach yielded patterns quite similar to the probability-based approach. The second supplemental analysis was a latent class analysis. Screening out youths predicted to be mischievous responders led to disparity estimates consistent with screening out the top 1 to 2% of extreme responders on the probability-based index. Thus, with the current dataset, the more complex models yielded patterns consistent with the simpler approaches.

[3] I did something similar for this study. Because several studies using anonymous SAQ data have found a correlation between LGBQ identification and weapon carrying or being overweight (e.g., Centers for Disease Control and Prevention, 2011; Garofalo et al., 1998, Grant et al., in press), I performed supplemental analyses with a pared-down screener consisting of *only* the items concerning deafness, blindness, and extreme height (because there is no known literature suggesting that such responses are related to LGBQ or transgender identification). The patterns from the sensitivity analysis using the pared-down screener were similar to those using the full screener, suggesting that (in this case) even a limited screener used for sensitivity analyses can help assess robustness and likely identifies the most mischievous responders.

REFERENCES

Anderson, K. (n.d.). *Gay agenda in school—A Christian worldview perspective*. Retrieved from http://www.probe.org/site/c.fdKEIMNsEoG/b.4219121/k.48F3/Gay_Agenda_in_Schools.htm

Anderson, K. (2008). *The Biblical point of view on homosexuality*. Eugene, OR: Harvest House Publishers.

Badgett, M. V. L. (2009). *Best practices for asking questions about sexual orientation on surveys*. Los Angeles: The Williams Institute, UCLA. Retrieved from http://escholarship.org/uc/item/706057d5

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1982. doi: 10.1037/0022-3514.51.6.1173

Blum, R. W., Kelly, A., & Ireland, M. (2001). Health-risk behaviors and protective factors among adolescents with mobility impairments and learning and emotional disabilities. *Journal of Adolescent Health*, *28*, 481–490. doi: 10.1016/S1054-139X(01)00201-4

Bontempo, D. E., & D'Augelli, A. R. (2002). Effects of at-school victimization and sexual orientation on lesbian, gay, or bisexual youths' health risk behavior. *Journal of Adolescent Health*, *30*, 364–374. doi: 10.1016/S1054-139X(01)00415-3

Centers for Disease Control and Prevention. (2011). *Sexual identity, sex of sexual contacts, and health-risk behaviors among students in grades 9-12: Youth risk behavior surveillance, selected cities, United States, 2001-2009* (MMWR Early Release 60). Atlanta, GA: U.S. Department of Health and Human Services.

Cheng, M. M., & Udry, J. R. (2002). Sexual behaviors of physically disabled adolescents in the United States. *Journal of Adolescent Health*, *31*, 48–58. doi: 10.1016/S1054-139X(01)00400-1

Coffman, K. B., Coffman, L. C., & Ericson, K. M. M. (2013). *The size of the LGBT population and the magnitude of anti-gay sentiment are substantially underestimated* (No. w19508). Cambridge, MA: National Bureau of Economic Research.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cornell, D., Klein, J., Konold, T., & Huang, F. (2012). Effects of validity screening items on adolescent survey data. *Psychological Assessment*, *24*, 21–35. doi: 10.1037/a0024824

D'Augelli, A. R., & Grossman, A. H. (2006). Researching lesbian, gay, and bisexual youth: Conceptual, practical, and ethical considerations. *Journal of Gay and Lesbian Issues in Education*, *3*, 35–56. doi: 10.1300/J367v03n02_03

Espelage, D. L., Aragon, S. R., Birkett, M., & Koenig, B. (2008). Homophobic teasing, psychological outcomes, and sexual orientation among high school students: What influence do parents and schools have? *School Psychology Review*, *37*(2), 202–216.

Fan, X. (2003). Correction to data and conclusions. *Child Development*, *74*, 65. doi: 10.1111/1467-8624.00579

Fan, X., Miller, B. C., Christensen, M., Park, K.- E., Grotevant, H. D., van Dulmen, M., et al. (2002). Questionnaire and interview inconsistencies exaggerated differences between adopted and non-adopted adolescents in a national sample. *Adoption Quarterly*, *6*, 7–27. doi: 10.1300/J145v06n02_02

Fan, X., Miller, B. C., Park, K., Winward, B. W., Christensen, M, Grotevant, H. D., et al. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods*, *18*, 223–244. doi: 10.1177/152822X06289161

Fanelli, D., & Ioannidis, J. P. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences*, *110*, 15031–15036. doi: 10.1073/pnas.1302997110

Faulkner, A. H., & Cranston, K. (1998). Correlates of same-sex sexual behavior in a random sample of Massachusetts high school students. *American Journal of Public Health*, *88*, 262–266. doi: 10.2105/AJPH.88.2.262

Garofalo, R., Wolf, R. C., Kessel, S., Palfrey, J., & DuRant, R. H. (1998). The association between health risk behaviors and sexual orientation among a school-based sample of adolescents. *Pediatrics*, *101*, 895–902. doi: 10.1542/peds.101.5.895

Gelman, A. (in press). The connection between varying treatment effects and the current crisis of unreplicable research in social science. *Journal of Management*.

Glenn, G. (2002). *Even Rosie knows homosexual adoption puts children at risk*. Retrieved from http://www.freerepublic.com/focus/news/648359/posts

Glynn, A. N. (2013). What can we learn with statistical truth serum? Design and analysis of the list experiment. *Public Opinion Quarterly*, *77*(S1), 159–172. doi: 10.1093/poq/nfs070

Grant, J. E., Odlaug, B. L., Derbyshire, K., Schreiber, L. R., Lust, K., & Christenson, G. (in press). Mental health and clinical correlates in lesbian, gay, bisexual, and queer young adults. *Journal of American College Health*. doi: 10.1080/07448481.2013.844697

Harris, K. M., Gordon-Larsen, P., Chantala, K., & Udry, J. R. (2006). Longitudinal trends in race/ethnic disparities in leading health indicators from adolescence to young adulthood. *Archives of Pediatrics & Adolescent Medicine*, *160*, 74–81. doi: 10.1001/archpedi.160.1.74

Hatzenbuehler, M. L. (2009). How does sexual minority stigma "get under the skin"? A psychological mediation framework. *Psychological Bulletin*, *135*, 707–730. doi: 10.1037/a0016441

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*(4), 765–789. doi: 10.1017/S0003055411000414

Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, *294*, 218–228. doi: 10.1001/jama.294.2.218

Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLoS Medicine*, *2*, e124. doi: 10.1371/journal.pmed.0020124

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865. doi: 10.1093/poq/nfn063

LeVasseur, M. T., Kelvin, E. A., & Grosskopf, N. A. (2013). Intersecting identities and the association between bullying and suicide attempt among New York City youths: Results from the 2009 New York City Youth Risk Behavior Survey. *American Journal of Public Health*, *103*, 1082–1089. doi: 10.2105/AJPH.2012.300994

Marshal, M. P., Friedman, M. S., Stall, R., King, K. M., Miles, J., Gold, M. A., . . . & Morse, J. Q. (2008). Sexual orientation and adolescent substance use: A meta-analysis and methodological review. *Addiction, 103*(4), 546–556. doi:10.1111/j.1360-0443.2008.02149.x

McCormack, M. (2011). The declining significance of homohysteria for male students in three sixth forms in the south of England. *British Educational Research Journal*, *37*(2), 337–353. doi: 10.1080/01411921003653357

McRee, A. L., Haydon, A. A., & Halpern, C. T. (2010). Reproductive health of young adults with physical disabilities in the US. *Preventive Medicine*, *51*, 502–504. doi: 10.1016/j.ypmed.2010.09.006

Meyer, I. H. (2003). Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: Conceptual issues and research evidence. *Psychological Bulletin*, *129*, 674–697. doi: 10.1037/0033-2909.129.5.674

Miller, B. C., Fan, X., Christensen, M., Grotevant, H. D., & Van Dulmen, M. (2000). Comparisons of adopted and nonadopted adolescents in a large, nationally representative sample. *Child Development*, *71*, 1458–1473. doi: 10.1111/1467-8624.00239

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*, 67–82. doi: 10.1093/esr/jcp006

Moss, P. A., Phillips, D. C., Erickson, F. D., Floden, R. E., Lather, P. A., & Schneider, B. L. (2009). Learning from our differences: A dialogue across perspectives on quality in education research. *Educational Researcher*, *38*(7), 501–517. doi: 10.3102/0013189X09348351

Poteat, V. P., Mereish, E. H., DiGiovanni, C. D., & Koenig, B. W. (2011). The effects of general and homophobic victimization on adolescents' psychosocial and educational concerns: The importance of intersecting identities and parent support. *Journal of Counseling Psychology*, *58*, 597–609. doi: 10.1037/a0025095

Poulin, C., MacNeil, P., & Mitic, W. (1993). The validity of a province-wide student drug use survey: Lessons in design. *Canadian Journal of Public Health*, *84*, 259–264.

Robinson, J. P., & Espelage, D. L. (2011). Inequities in educational and psychological outcomes between LGBTQ and straight students in middle and high school. *Educational Researcher*, *40*, 315–330. doi: 10.3102/0013189X11422112

Robinson, J. P., & Espelage, D. L. (2012). Bullying explains only part of LGBTQ–heterosexual risk disparities: Implication for policy and practice. *Educational Researcher*, *41*, 309–319. doi: 10.3102/0013189X12457023

Robinson, J. P., & Espelage, D. L. (2013). Peer victimization and sexual risk differences between lesbian, gay, bisexual, transgender or questioning and nontransgender heterosexual youths in grades 7-12. *American Journal of Public Health*, *103*(10), 1810–1819. doi: 10.2105/AJPH.2013.301387

Russell, S. T. (2005). Beyond risk: Resilience in the lives of sexual minority youth. *Journal of Gay and Lesbian Issues in Education*, *2*, 5–18. doi: 10.1300/J367v02n03_02

Russell, S. T., & Joyner, K. (2001). Adolescent sexual orientation and suicide risk: Evidence from a national study. *American Journal of Public Health*, *91*, 1276–1281. doi: 10.2105/AJPH.91.8.1276

Russell, S. T., Sinclair, K. O., Poteat, V. P., & Koenig, B. W. (2012). Adolescent health and harassment based on discriminatory bias. *American Journal of Public Health*, *102*, 493–495. doi: 10.2105/AJPH.2011.300430

Saewyc, E. M. (2011). Research on adolescent sexual orientation: Development, health disparities, stigma, and resilience. *Journal of Research on Adolescence*, *21*, 256–272. doi: 10.1111/j.1532-7795.2010.00727.x

Saewyc, E. M., Bauer, G. R., Skay, C. L., Bearinger, L. H., Resnick, M. D, Reis, E., & Murphy, A. (2004). Measuring sexual orientation in adolescent health surveys: Evaluation of eight school-based surveys. *Journal of Adolescent Health*, *35*, e1–e16. doi: 10.1016/j.jadohealth.2004.06.002

Saewyc, E. M., Skay, C. L., Hynds, P., Pettingell, S., Bearinger, L. H., Resnick, M. D., & Reis, E. (2008). Suicidal ideation and attempts among adolescents in North American school-based surveys: Are bisexual youth at increasing risk? *Journal of LGBT Health Research*, *3*(2), 25–36. doi: 10.1300/J463v03n02_04

Saewyc, E., Skay, C., Richens, K., Reis, E., Poon, C., & Murphy, A. (2006). Sexual orientation, sexual abuse, and HIV-risk behaviors among adolescents in the Pacific Northwest. *American Journal of Public Health*, *96*(6), 1104–1110. doi: 10.2105/AJPH.2005.065870

Savin-Williams, R. C. (2001). A critique of research on sexual-minority youths. *Journal of Research on Adolescence*, *24*, 5–13. doi: 10.1006/jado.2000.0369

Savin-Williams, R. C. (2005). *The new gay teenager*. Cambridge, MA: Harvard University Press.

Savin-Williams, R. C., & Joyner, K. (in press). The dubious assessment of gay, lesbian, and bisexual adolescents of Add Health. *Archives of Sexual Behavior*. doi: 10.1007/s10508-013-0219-5

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*, 859–883. doi: 10.1037/0033-2909.133.5.859

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science*, *280*, 867–873. doi: 10.1126/science.280.5365.867

Wolter, F., & Preisendörfer, P. (2013). Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods & Research*, *42*(3), 321–353. doi: 10.1177/0049124113500474

AUTHOR

JOSEPH P. ROBINSON-CIMPIAN, PhD, is an assistant professor of quantitative and evaluative research methodologies in the Department of Educational Psychology at the University of Illinois at Urbana-Champaign, 210F Education Bldg., 1310 S. 6th St., Champaign, IL 61820; *jpr@illinois.edu*. His research focuses on the use of novel and rigorous methods to study equity and policy, particularly concerning sexual minorities, women, and language minorities.

Manuscript received October 28, 2013
Revisions received January 24, 2014, and March 5, 2014
Accepted April 10, 2014

Appendix A

Constructing P*, a Continuous Measure of Extreme-Response Patterns*

In general terms, individual *i*'s value of *P* is the (weighted) product of *i*'s response probabilities *p* for each item *m* in a group of items *M*, and each probability is weighted by item-specific weight *w*:

$$P_i = \prod_{m=1}^{M} \left(p_{im}\right)^{w_m}.$$

Although p_{im} can be any kind of probability (e.g., a conditional probability), in this article p_{im} is just a simple probability (i.e., the proportion of individuals who provided the response that individual i provided for item m). For example, if 10% of the responses were 1, and 90% were 2, then an individual who provided a response of 1 would have $p_{im} = .1$, and an individual who provided a response of 2 would have $p_{im} = .9$ for that item.

Regarding the weights (the w_m s), one may want to weight some items more heavily than others (as was done with approach *P3*, one of two additional probability-based screener-indexing approaches developed for this article, described in Appendix Table A1), and this general formula can accommodate this. For example, if one item had two responses with probabilities $p = \{.10, .90\}$ and a different item had 10 responses with probabilities $p = \{.09, .11, .1, \ldots, .1\}$, one might want to give additional weight to the first item because it has a clear low-probability response option relative to the other options. In such a case, w could incorporate the representativeness of the response options in a manner similar to the multigroup entropy index Theil's *H* (Reardon & Firebaugh, 2002). Higher values of *H* indicate greater equality of responses (e.g., the second item), and lower values of *H* indicate lower equality of responses (e.g., the first item). Thus, if one wishes to give additional weight to the items with clear low-response options, one need simply take the inverse of *H*—that is, $w_m = H_m^{-1}$. This approach was used for *P3* in this article.

In approach *P1*, each item *m* is dichotomous; this facilitates comparisons with the count-based screener (*C1*). But note that *P* does not require dichotomous items, a feature that *P2* and *P3* capitalize on (see Appendix Table A1). For instance, for the item "When was the last time you were seen by a dentist?" there were four possible response options, and each has its own p_r. Thus, each individual will have $p_{im} = p_{rm}$ when i chooses $R = r$. In approaches *C1* and *P1*, the three categories with the highest responses were collapsed into one category, thereby dichotomizing the variable into the single lowest frequency response option and all other options. But in approaches *P2* and *P3*, the original polytomous response options were preserved intact for aggregation.

In principle, *P* can be a reflection of as many items *M* as the researcher desires and has access to, taking full advantage of all possible response combinations; however, in practice, one should select the items judiciously and so as to reduce the likelihood of falsely identifying mischievous responders. For instance, if examining LGBQ–heterosexual disparities, it would be unwise to include among the *M* items an item on attraction to the same sex. Although youths may find response options for this item funny, those responses are in principle related to LGBQ identification. Thus, there is a delicate balance between identifying items (and particular response options) that mischievous responders may find alluring and not including items that are likely correlated with group identification.

REFERENCE

Reardon S. F., & Firebaugh G. (2002). Measures of multigroup segregation. *Sociological Methodology, 32*(1), 33–67.

Table A1
Existing-Item Screener-Indexing Approaches Used for the Sensitivity Analyses

Method	General Approach	How Used in This Study / Comments
C1: Count-based low-frequency response-probabilities (LFRPs)	Count the number of low-frequency responses provided on the set of screener items, and then estimate the disparities after removing individuals who provide more than a given number of low-frequency responses to the screener items.	Robinson and Espelage (2011) originally proposed this approach, where (after excluding respondents who provided extreme height/weight values) they excluded any respondent who provided two or more low-frequency responses to eight screener items (which differ somewhat from the items in the current article due to different questionnaires used). But we can look more generally at how the disparity estimates change by using different count thresholds (e.g., excluding only those who provide four or more low-frequency responses).
P1: Percentiles of a continuous measure of extreme-response patterns (*P*), constructed from LFRPs	Creates a composite measure of extreme responses to screener items by multiplying the probabilities of the dichotomized items together, thereby implicitly providing additional weight to items with *lower* low-frequency response options when identifying potentially mischievous responders. Values of *P1* are then ranked.	Once ranked along *P1*, we can examine (1) the proportions of reported LGBQ- or transgender-identified youths or youths reporting physical disabilities at various percentiles of *P1* and (2) how the estimated disparities change as observations below a given percentile of *P1* are removed.
P2: A version of *P* constructed from all response-option probabilities from the screener items	*P1* used the probabilities from items that have been dichotomized to indicate whether the low-frequency response was chosen. But that approach may mask relevant information that was contained in the original polytomous items before they were dichotomized. *P2* multiplies the *original* item response-option probabilities together.	*P2* would allow one to distinguish a respondent who replied he had not seen a dentist in 2 years from one who said he had seen the dentist within the past year, whereas all above approaches mask this distinction. *P2* also does not require the researcher to predefine unusual response options that may be "funny," but rather allows the data analysis to consider all response options, which may reveal different patterns.
P3: A version of *P* that uses all response-option probabilities from the screener items and weights them by the variation between items	This approach is similar to *P2* but applies additional weight to items that have more uneven distributions in the response options, in an attempt to provide additional weight to items that may be more helpful in identifying mischievous responders.	Some items in the screener display more uniformity in response-option frequency, whereas other items have less uniformity. For example, students may respond in nearly equal proportions to the non-"funny" options for some items. By contrast, other items may have one nonfunny response that is selected far more often than others. Thus, *P3* would weight the item with the dominant nonfunny response more heavily because it has a more uneven distribution of responses, and therefore may be more likely to reveal responders who are drawn to items with response options that are more clearly unusual.

Note. The count-based approach (*C1*) and the first probability-based approach (*P1*) are discussed in the main text. Approaches *P2* and *P3* appear only here, for brevity in the main text and because similar patterns were seen across the different approaches.

Table A2
Regression-Based Estimated Disparities, by Outcome, Group, and Screener-Indexing Approach Used as Covariates

			Probability-Based as Covariate		
	No Additional Covariates	Count-Based as Covariate	Dichotomous (Shown in Main Text)	Polytomous (Introduced in Appendix A)	Polytomous, Weighted (Introduced in Appendix A)
Suicidal ideation					
LGBQ–heterosexual	0.740 (0.111), $p < 0.001$	0.232 (0.050), $p < 0.001$	0.236 (0.051), $p < 0.001$	0.231 (0.055), $p < 0.001$	0.274 (0.059), $p < 0.001$
LG–heterosexual	1.266 (0.332), $p = 0.001$	0.342 (0.096), $p = 0.002$	0.364 (0.123), $p = 0.007$	0.355 (0.149), $p = 0.026$	0.509 (0.160), $p = 0.005$
B–heterosexual	0.496 (0.143), $p = 0.002$	0.262 (0.108), $p = 0.025$	0.256 (0.102), $p = 0.020$	0.244 (0.111), $p = 0.039$	0.250 (0.115), $p = 0.041$
Q–heterosexual	1.119 (0.269), $p < 0.001$	0.218 (0.091), $p = 0.026$	0.234 (0.099), $p = 0.027$	0.267 (0.112), $p = 0.027$	0.342 (0.131), $p = 0.017$
Transgender–cisgender	1.985 (0.318), $p < 0.001$	0.150 (0.138), $p = 0.289$	0.153 (0.158), $p = 0.342$	0.254 (0.156), $p = 0.118$	0.439 (0.186), $p = 0.028$
Disabled–nondisabled	1.877 (0.228), $p < 0.001$	0.641 (0.120), $p < 0.001$	0.706 (0.111), $p < 0.001$	0.978 (0.144), $p < 0.001$	1.027 (0.150), $p < 0.001$
School belongingness					
LGBQ–heterosexual	-0.525 (0.078), $p < 0.001$	-0.294 (0.056), $p < 0.001$	-0.294 (0.057), $p < 0.001$	-0.243 (0.051), $p < 0.001$	-0.269 (0.055), $p < 0.001$
LG–heterosexual	-0.751 (0.194), $p = 0.001$	-0.377 (0.150), $p = 0.020$	-0.379 (0.147), $p = 0.017$	-0.362 (0.131), $p = 0.012$	-0.412 (0.134), $p = 0.006$
B–heterosexual	-0.409 (0.080), $p < 0.001$	-0.291 (0.072), $p = 0.001$	-0.295 (0.072), $p = 0.001$	-0.211 (0.071), $p = 0.007$	-0.233 (0.073), $p = 0.004$
Q–heterosexual	-0.617 (0.103), $p < 0.001$	-0.273 (0.077), $p = 0.002$	-0.273 (0.074), $p = 0.001$	-0.265 (0.068), $p = 0.001$	-0.282 (0.080), $p = 0.002$
Transgender–cisgender	-1.015 (0.168), $p < 0.001$	-0.225 (0.141), $p = 0.125$	-0.225 (0.146), $p = 0.138$	-0.223 (0.139), $p = 0.122$	-0.295 (0.149), $p = 0.062$
Disabled–nondisabled	-0.750 (0.132), $p < 0.001$	-0.176 (0.089), $p = 0.060$	-0.210 (0.087), $p = 0.025$	-0.304 (0.102), $p = 0.007$	-0.289 (0.106), $p = 0.012$
Cocaine/crack use					
LGBQ–heterosexual	0.750 (0.110), $p < 0.001$	0.242 (0.064), $p = 0.001$	0.252 (0.067), $p = 0.001$	0.197 (0.063), $p = 0.005$	0.258 (0.066), $p = 0.001$
LG–heterosexual	1.038 (0.337), $p = 0.006$	0.185 (0.177), $p = 0.310$	0.196 (0.190), $p = 0.314$	0.111 (0.189), $p = 0.564$	0.282 (0.212), $p = 0.199$
B–heterosexual	0.521 (0.122), $p < 0.001$	0.295 (0.085), $p = 0.002$	0.299 (0.087), $p = 0.002$	0.210 (0.088), $p = 0.026$	0.226 (0.088), $p = 0.017$
Q–heterosexual	1.141 (0.148), $p < 0.001$	0.273 (0.120), $p = 0.034$	0.295 (0.122), $p = 0.024$	0.303 (0.112), $p = 0.014$	0.395 (0.124), $p = 0.004$
Transgender–cisgender	2.322 (0.137), $p < 0.001$	0.491 (0.147), $p = 0.003$	0.518 (0.148), $p = 0.002$	0.562 (0.137), $p = 0.001$	0.775 (0.148), $p < 0.001$
Disabled–nondisabled	1.762 (0.188), $p < 0.001$	0.508 (0.106), $p < 0.001$	0.512 (0.114), $p < 0.001$	0.773 (0.125), $p < 0.001$	0.774 (0.119), $p < 0.001$

Note. Cluster-robust standard errors appear in parentheses. Each regression for column “No additional covariates” contains only the LGBQ (or subgroup; e.g., LG), transgender, or disability indicator variable as a covariate. Each regression for “Count-based as covariate” contains the group indicator and an indicator for each number of low-frequency responses. Each regression for “Probability-based as covariate” contains the group indicator and a fourth-order polynomial for the continuous function of the natural log of *P*.