AERA Open
October-December 2019, Vol. 5, No. 4, pp. 1–35
DOI: 10.1177/2332858419888892
Article reuse guidelines: sagepub.com/journals-permissions
© The Author(s) 2019. http://journals.sagepub.com/home/ero

Large-Scale Estimates of LGBQ-Heterosexual Disparities in the Presence of Potentially Mischievous Responders: A Preregistered Replication and Comparison of Methods

Joseph R. Cimpian
New York University
Jennifer D. Timmer
Vanderbilt University

Although numerous survey-based studies have found that students who identify as lesbian, gay, bisexual, or questioning (LGBQ) have elevated risk for many negative academic, disciplinary, psychological, and health outcomes, the validity of the types of data on which these results rest have come under increased scrutiny. Over the past several years, a variety of data-validity screening techniques have been used in attempts to scrub data sets of "mischievous responders," youth who systematically provide extreme and untrue responses to outcome items and who tend to falsely report being LGBQ. We conducted a preregistered replication of Cimpian et al. with the 2017 Youth Risk Behavior Survey to (1) estimate new LGBQ-heterosexual disparities on 20 outcomes; (2) test a broader, mechanistic theory relating mischievousness effects with a feature of items (i.e., item response-option extremity); and (3) compare four techniques used to address mischievous responders. Our results are consistent with Cimpian et al.'s findings that potentially mischievous responders inflate LGBQ-heterosexual disparities, do so more among boys than girls, and affect outcomes differentially. For example, we find that removing students suspected of being mischievous responders can cut male LGBQ-heterosexual disparities in half overall and can completely or mostly eliminate disparities in outcomes including fighting at school, driving drunk, and using cocaine, heroin, and ecstasy. Methodologically, we find that some methods are better than others at addressing the issue of data integrity, with boosted regressions coupled with data removal leading to potentially very large decreases in the estimates of LGBQ-heterosexual disparities, but regression adjustment having almost no effect. While the empirical focus of this article is on LGBQ youth, the issues discussed are relevant to research on other minority groups and youth generally, and speak to survey development, methodology, and the robustness and transparency of research.

Keywords: *at-risk students, descriptive analysis, gay/lesbian studies, mischievous responders, replication, research methodology, secondary data, survey research, validity/reliability*

While much survey-based research suggests that students who identify as lesbian, gay, bisexual, or questioning (LGBQ) are at elevated risk for a wide variety of negative outcomes—including suicide attempts, drug and alcohol use, sexual risk taking, being bullied, and facing disciplinary action (e.g., Centers for Disease Control and Prevention [CDC], 2016; Espelage, Aragon, Birkett, & Koenig, 2008; Mittleman, 2018; Robinson & Espelage, 2011; Russell, Sinclair, Poteat, & Koenig, 2012; Saewyc et al., 2004)—recent research has called into question the validity of the data on which many of these claims are based (e.g., Cimpian, 2017; Cimpian et al., 2018; Robinson-Cimpian, 2014; cf. Savin-Williams & Joyner, 2014a, 2014b; but cf. Fish & Russell, 2018; Katz-Wise, Calzo, Li, & Pollitt, 2015; Li, Katz-Wise, & Calzo, 2014). More specifically, the research challenging the validity of the data argues that some of the youth completing the surveys may have been "jokesters" or "mischievous responders" who provided dubious responses possibly because they found it funny to claim they were not heterosexual on a survey and also to make other bogus claims about their risk and misconduct (Cimpian et al., 2018; Robinson-Cimpian, 2014; cf. Savin-Williams & Joyner, 2014a, 2014b). For example, a student who identifies as heterosexual and does not use drugs may find it amusing to report on a survey that he identifies as gay and uses drugs often and heavily, thereby inflating estimates of gay-identified youth using drugs (see Fan et al., 2006, for similar argumentation and evidence with respect to adopted, foreign-born, and disabled statuses). Such claims of invalid survey data from youth leading to elevated risk profiles are not new (see, e.g., Cornell, Klein, Konold, & Huang, 2012; Cornell & Loper, 1998; Cross & Newman-Gonchar, 2004; Fan et al., 2006; Furlong, Sharkey, Bates, & Smith, 2004; Rosenblatt & Furlong, 1997); however, there is a new and increasing emphasis on how data invalidity can differentially affect estimates of minority-group risk, particularly when it is challenging or impossible to verify responses, such as in the case of LGBQ identification. Because of their systematic patterns of extreme reporting and their propensity to (falsely) report minority-group membership, the bias introduced into estimates by mischievous responders is distinctly different from other forms of misreporting bias such as haphazard responding and misunderstanding of terminology, and can exert extreme bias into estimates of minority-group well-being (Cimpian, 2017; Fan et al., 2006; cf. Groves, Fowler, Couper, Lepkowski, & Tourangeau, 2011).

It is imperative to understand and appropriately address mischievous responders for at least three reasons, all of which we discuss in this article: (1) Mischievous responders lead to incorrect estimates of the risk of minority groups (e.g., LGBQ youth, transgender youth, racial/ethnic minorities, students with disabilities) and impede our understanding of how to improve outcomes for these subgroups. (2) Survey designers need to know what types of items are particularly susceptible to mischievous responding. (3) Methodologists (and applied researchers) need to know the best ways to detect and reduce the effects of mischievous responding. Thus, the study of mischievous responders cuts across many aspects of education and social science research. This article will be of interest to researchers of LGBQ youth, the subject of our empirical investigation and of much of the current debates around mischievous responders. But it will also be of interest to survey designers and methodologists, as well as to researchers of other subgroups or of youth well-being in general.

In this article, we expand the field's understanding of the effects of potentially mischievous responders in three ways: First, we perform a *preregistered replication* of Cimpian et al. (2018) with a recently released data set collected by the CDC, which informs how potentially mischievous responders affect estimates of LGBQ-heterosexual disparities on 20 commonly examined outcomes in a data set containing $N = 108{,}093$ student survey records in its final analytic form. We hypothesize that the removal of potentially mischievous responders will lead to significant reductions in disparities, on average, as soon as 1% of observations are removed, with larger reductions for males than females. Second, we replicate Cimpian et al.'s (2018) analysis exploring the relationship between item response-option extremity and the effects of mischievous responders, providing new large-scale evidence to a broader theory about the *types* of items and their response options that are most likely to be influenced by mischievous responders—this has implications well beyond LGBQ-heterosexual disparities, and can also help researchers think about how to address mischievous responding in the *early* stages of research. We hypothesize that disparities for items with relatively fewer respondents choosing the most extreme option (e.g., selecting "40+ days" for heroin use) will be more affected by screening out likely mischievous responders than items with more frequently selected extreme options (e.g., reporting feeling sad). This is because, having selected the extreme option for both types of items, mischievous responders will make up a disproportionately large share of extreme response selectors for items of the former type. Third, we provide a *direct empirical comparison of the methods* proposed—and implemented in the literature—to address potentially mischievous responders in a single data set, thereby eliminating one source of variability in comparing effects across different published papers (and their different data sets used), as well as providing insights to the field on when different methods reach similar conclusions about disparities. This component of the study is more exploratory, and we therefore make no predictions regarding methodological differences. Finally, we provide some guidance to researchers on how to *combine preregistration practices and sensitivity analyses* to improve the transparency of addressing data-validity threats from mischievous responders.

The Importance of Data Validity for LGBQ Research in Education

Research on LGBQ (and more broadly, LGBTQ) youth in education has been receiving increasing focus as of late, with a recent American Educational Research Association–published book (Wimberly, 2015b) and special issue of *Educational Researcher* (Cimpian & Herrington, 2017), as just two illustrations. Although both the book and special issue include a wide variety of research methodological perspectives, the general trend in education research on LGBQ youth has been a movement toward quantitative work (Brockenbrough, 2017). For instance, Wimberly (2015a, 2015b; Wimberly & Battle, 2015) calls for more quantitative research on LGBQ students. Thus, it is noteworthy that there have been several recent critiques from across methodological perspectives regarding how quantitative research chooses to categorize LGBQ youth in education research, yet it is valued by the education research community (Brockenbrough, 2017; Cimpian, 2017; Mayo, 2017). Indeed, if we are experiencing an increasing quantification of LGBQ research in education, we must ensure that the research is valid and not valued simply because it is quantitative. As seen in many research studies (for critiques, see Gelman & Loken, 2014; Simmons, Nelson, & Simonsohn, 2011), quantitative education researchers of LGBQ youth make many choices throughout the research process that can affect the findings. Because the presence of mischievous responders calls into question the validity of the data as well as any associated findings, we hope to draw attention to the choices regarding data validity and mischievous responders, as well as illustrate how these choices affect outcomes, while providing guidance on how to make the research choices more transparent.

It is also worth noting that, in an era of growing importance of administrative data sets for education research (Dynarski & Berends, 2015), LGBQ status is distinctly different from variables like race/ethnicity, sex, special education status, and English learner status because LGBQ status is *not* collected in administrative data sets. Thus, large-scale quantitative education research on LGBQ youth relies almost exclusively on the type of data gathered through anonymous self-administered questionnaires such as the one that is the focus of this article.[1]

Screening, Its Different Purposes, and Assumptions Required

Before elaborating on the specific contributions of this article, it is important to note the broad assumptions required of any analysis with self-administered questionnaire data, and also to distinguish the general assumptions of the techniques we will discuss here from those of other data-validity methods in the literature. In doing so, we also hope to clarify why we are focusing on this specific set of data-validity sensitivity analyses.

Implicit in any data analysis with self-administered questionnaire data are assumptions about the validity of the data. Retaining all of the data, as many researchers do, assumes that the data are valid as they are. Removing questionable observations also makes assumptions, which vary by the method used and the nature of the analysis. Some data-removal techniques focus on assessing the variability in survey responses (e.g., Cross & Newman-Gonchar, 2004; Meade & Craig, 2012; Shukla & Konold, 2018). For example, one way Shukla and Konold (2018) identified suspect responses was by examining the variability within individuals in their responses to items that are part of a common construct scale—that is, we would expect a relatively small range of variability within an individual when responding to items that tap into the same construct—and then using latent profile analysis to identify and remove individuals who exhibited high levels of response inconsistency across seven different constructs. A method such as this identifies cases on their responses to *outcome* items of interest (e.g., academic press), whereas the typical screening responses we will focus on identify unusual cases based on responses to items that are *not* outcomes of interest (e.g., height). Other data-removal methods ask respondents to rate how truthful they were in their responses (e.g., Cornell et al., 2012; Jia, Konold, Cornell, & Huang, 2018; Shukla & Konold, 2018), thereby making the assumption that the respondents will truthfully state how untruthful they were earlier in the survey; however, it should be noted that reported truthfulness does correlate with more complex methods of detecting response-inconsistent data (Shukla & Konold, 2018).

These examples illustrate some dimensions on which the screeners we examine differ from some other approaches, but there are also important differences in their intent, which correspond to different required assumptions. For instance, the intent of many researchers interested in removing invalid data is to obtain a more accurate measure of an outcome *globally*, that is, across all groups in the data (e.g., Cornell et al., 2012; Furlong et al., 2017; Jia et al., 2018; Meade & Craig, 2012; Shukla & Konold, 2018). When researchers seek to obtain a global estimate (e.g., bullying experienced by all students), then the assumption when screening out observations is that the screener itself does not introduce bias into the global estimate—in other words, the assumption that screening has *no global impact* that would introduce bias into the estimate of the true overall value. By contrast, researchers interested in comparing groups (e.g., differences between LGBQ- and heterosexual-identified students in reported experiences of bullying) need only make the *weaker* assumption of *no differential impact* of screening on the groups being compared. This assumption requires that the screening technique does not introduce bias into the estimate for one group differently than for the group to which it is compared.

Thus, the assumptions required for data removal are *weakened* when making comparisons across groups, making data-removal in the case of comparisons less assumption-laden (though there are still assumptions, which we will discuss in detail later). The tradeoff, however, is that the assumptions required for valid data-removal in the case of one group comparison (e.g., LGBQ-heterosexual disparities) may not be plausibly satisfied for a different group comparison (e.g., disabled–non-disabled disparities) and may not be plausible more globally. For instance, Robinson-Cimpian (2014) assumed that actual LGBQ- and heterosexual-identified youth should not differ in terms of reporting blindness or deafness, and so he included those items in his screener when estimating LGBQ-heterosexual disparities; however, those same items of blindness and deafness *are* expected to differ between disabled and non-disabled students, and so those items were *excluded* from his screener when estimating disabled–non-disabled disparities because their inclusion would render implausible the assumption of non-differential impact.

Because the data-validity methodology literature is expanding and producing a large set of techniques (see, e.g., Fan et al., 2006; Jia et al., 2018; Shukla & Konold, 2018), it is necessary for us to hone the focus of this article on the most relevant techniques for the topic under study: LGBQ-heterosexual disparities. As such, we will focus our attention on this latter group of data-validity techniques, those which are primarily used to compare groups, and especially used to compare LGBQ and heterosexual youth (e.g., Cimpian et al., 2018; Fish & Russell, 2018; Mittleman, 2018; Robinson-Cimpian, 2014). We also focus on cases where data are anonymized, a common "best practice" when gathering sensitive data such as sexual identity (Badgett, 2009; Tourangeau &

Yan, 2007), but one that prohibits verification through triangulation (Fan et al., 2006). These cases are perhaps the most methodologically challenging, require more assumptions, and arguably could benefit the most from replication and a comparison of existing methods.

The Present Article

As alluded to above, the current article has three components (separated as Studies). Study 1 is a direct replication of Cimpian et al. (2018) examining how potentially mischievous responders may affect estimates of LGBQ-heterosexual disparities. Study 2 further replicates Cimpian et al. (2018) by considering whether item-response extremity plays a role in those effects. Finally, in Study 3, we compare several common approaches for identifying and removing potentially mischievous responders. Here, we discuss each study in more detail.

First, replication is essential to ensuring that findings of a single study are not anomalous. Education researchers, and indeed the broader field of social scientists, are under increasing pressure to replicate and preregister studies to ensure the robustness of published findings (Fanelli & Ioannidis, 2013; Gehlbach & Robinson, 2018; Makel & Plucker, 2014; Open Science Collaboration, 2015). In this article, we conduct a direct replication of a recent study by Cimpian et al. (2018). The study by Cimpian et al. is an ideal one to replicate as part of a preregistered replication because Cimpian and colleagues used a data set that is part of a biannually collected series conducted by the CDC, and the 2017 iteration of the survey data was not yet released by the CDC when we submitted this manuscript as a registered report for consideration with our detailed analysis plan—thus, allowing for a preregistered hypothesis-testing study using a recurring national data set (Gehlbach & Robinson, 2018). However, Cimpian et al.'s (2018) study is important to directly replicate with new data for other reasons: First, the series the data come from are national and publicly available and have had a tremendous impact on the fields of education, psychology, and health (for some recent examples, see CDC, 2016, 2017; Clayton, Lowry, August, & Jones, 2016; Raifman, Moscoe, Austin, & McConnell, 2017; Vagi, Olsen, Basile, & Vivolo-Kantor, 2015; Zaza, Kann, & Barrios, 2016). Second, Cimpian et al.'s (2018) findings that LGBQ-heterosexual disparities may be substantially overestimated need to be replicated, given the impact the data series has on several fields and the questions raised about its validity, as well as what the findings could mean to methodological practices in survey-based comparisons, especially along sexuality dimensions. Thus, Study 1 of the present article is a direct preregistered replication of Cimpian et al. (2018) with the 2017 Youth Risk Behavior Survey (YRBS; for the preregistered form, go to https://aspredicted.org/sz9aa.pdf).

Second, Study 2 is also a direct preregistered replication of a component of Cimpian et al. (2018), testing the heterogeneity of effects on outcomes as related to extreme response selection. Cimpian and colleagues concluded that potentially mischievous responders affected LGBQ-heterosexual disparities on *average*, but there was substantial *heterogeneity* in how much the 20 outcomes were affected. Because mischievous responders are expected to provide low-frequency, extreme responses (Fan et al., 2006; Furlong, Fullchange, & Dowdy, 2017; Furlong, Sharkey, Bates, & Smith, 2004; Robinson-Cimpian, 2014), Cimpian and colleagues hypothesized that outcome items containing response options that were less frequently chosen (e.g., using heroin "40 or more times" in one's life) would be the items most affected by the removal of potentially mischievous responders. Cimpian et al. (2018) found strong support for this hypothesis in the 2015 state and district sample, with large standardized *B*s of 0.75 (*p*s < .001), among both males and females. Replicating this finding with the 2017 YRBS has implications for providing preregistered empirical support for this theory on how mischievous responders affect outcome estimates. In doing so, the work also provides survey developers and researchers with useful information regarding how they can mitigate the effects of mischievous responders in the *early stages* of survey research (i.e., survey development) instead of in the later stages (i.e., data analysis) through item construction. Therefore, Study 2 replicates the Cimpian et al. analysis of how item response option extremity relates to the effects of screening on individual outcome items.

Third, as discussed above, data-validity sensitivity techniques are gaining popularity in the study of LGBQ youth (e.g., Cimpian et al., 2018; Fish & Russell, 2018; Mittleman, 2018); yet, we would be remiss to conclude that all techniques make the same assumptions, that they all lead to the same conclusions, or even that the same techniques are applied similarly across different research studies. Thus, in addition to replicating the recent large-scale CDC-based study by Cimpian et al. (2018) with newly released data in Studies 1 and 2, this article compares the effects of the main data-validity sensitivity techniques used in the growing literature on LGBQ-heterosexual disparities in Study 3; and, it does so within a single data set, thereby eliminating one source of variability across different efforts to identify and address likely invalid data.

Method

Data

For all three studies, we use the publicly available YRBS data from the CDC: https://www.cdc.gov/healthyyouth/data/yrbs/data.htm. The students completed the paper-and-pencil questionnaires in school and answered questions related to their mental, emotional, sexual, and physical health. Following the approach of Cimpian et al. (2018), the final analytic sample is restricted to observations from the State and District YRBS data set with valid sampling weights and nonmissing values for sex and sexual identity. The CDC does not require surveying agencies to include the item on sexual identity; therefore, some entire states and districts are necessarily omitted from the analytic sample, just as in Cimpian et al. (2018).

Because we are primarily interested in replicating the methods and general findings of the earlier study, we use all states and districts that included the necessary sexual identity item in the 2017 survey, regardless of whether they were part of the 2015 sample in Cimpian et al. (2018). We do not expect estimates in this replication to be exact; instead, we are interested in the general trends of how estimates are affected by the removal of potentially mischievous responders. Nonetheless, there is substantial overlap in the jurisdictions that participated in the 2015 and 2017 YRBS. The majority of jurisdictions in the 2015 sample also appear in the 2017 analysis (i.e., 27 of 36 jurisdictions from 2015 are in the 2017 study), with four jurisdictions dropping out and nine being added (see Table 1). The final analytic sample was 108,093 students, of which 52,753 reported being males (6,219 of them reported LGBQ identifications) and the remaining 55,340 reported being females (12,228 of whom reported LGBQ identification; see Table 2).

In addition to providing details on the overlap in the publicly available jurisdictions meeting the study criteria from the 2015 and 2017 YRBS, Table 1 also shows which jurisdictions asked which of the seven items used in our screener. Jurisdictions asking all seven items in the 2017 YRBS appear in boldface. This is important for our comparison of methods in Study 3—while the boosted regression approach can more flexibly handle missing data, the presence of completely missing item-level data for entire jurisdictions unnecessarily complicates any comparison of methods. Thus, for Study 3, we restrict our analyses to only the subsample of jurisdictions asking all seven screener items (which we term the "full-screener" sample in our tables). For consistency across Studies 1, 2, and 3, the main text and focus of the article will be on the full-screener sample; however, for completeness, we also estimated all of Studies 1 and 2 for the full sample, and present those results in Appendix B. The results are generally similar.

The smaller full-screener sample ($N = 51,524$)—which is the main analytic sample for the remainder of this article—is demographically similar to the full sample (see Table 2). Furthermore, Table 3 illustrates that removal of observations due to screening itself (discussed below) does not alter the demographics of the sample in any substantial way, suggesting good generalizability regardless of sample restrictions and screening.

Outcomes

The YRBS includes a variety of items asking about high school students' risk-taking behaviors and attitudes. Following Cimpian et al. (2018), for both Studies 1 and 2 (as well as for our comparison of methods in Study 3), we examine 20 items commonly studied in LGBQ research. Specifically, we include the following outcomes: rode in a car with a drunk driver, drove drunk, skipped school because felt unsafe, fought at school, was forced into sex, their partner forced sex on them, was bullied at school, felt sad/hopeless, considered suicide, planned suicide, attempted suicide, smoking, alcohol use, cocaine use, heroin use, ecstasy use, steroids use, number of sex partners, physical activity, and TV watching (see the Users Manual [CDC, 2018] for specific phrasing of items, also contained in Appendix C). All outcomes are coded continuously (e.g., reporting "20 to 39 times" is coded as 29.5).

Study 1: Direct Replication of Cimpian et al.'s (2018) Disparity Estimate Effects of Mischievous Responders

Identification of Potentially Mischievous Responders. Cimpian et al. (2018) extended the study of potentially mischievous responders to the largest sample to date, introduced the application of boosted regressions (a machine-learning technique) to identify unusual responding patterns, focused on LGBQ-heterosexual disparities, and found that potentially mischievous responders may account for an average of 46% of the LGBQ-heterosexual youth outcome disparity among males and 23% among females. We use their approach but applied to the 2017 version of the YRBS.

We identify potentially mischievous responders by exploiting relationships between reporting being LGBQ and ostensibly unrelated survey items. We expect no real relationship between sexuality and student characteristics such as height, asthma diagnosis, or dental history; likewise, the frequency with which individuals eat carrots, fruit, potatoes, or salad are not expected to be associated with sexuality in reality. However, some youth might find it funny to report extreme responses (Furlong et al., 2004; Furlong et al., 2017; Robinson-Cimpian, 2014), for example, reporting eating copious amounts of fruit, having never been to a dentist, being extremely tall, and being gay, even if all of these are untrue. Thus, the youth providing these mischievous responses create spurious relationships between the predictor (screener) items (e.g., salad consumption, asthma diagnosis) and sexuality, which can lead to unexpected and potentially misleading estimates of disparities. These screener items can then be used to identify youth providing the most unusual patterns of responses, and disparities in outcomes can be estimated without these potentially problematic responses included in the data.

As described in Cimpian et al. (2018), boosted regression (Friedman, 2001) is a machine-learning technique we can use to predict reporting LGBQ identification as a function of the specified screener survey items. We use the same seven screener items, location fixed effects to account for variation in survey item inclusion across jurisdictions, and YRBS survey weights (DuGoff, Schuler, & Stuart, 2014) as predictors.

TABLE 1

Screener Items Administered by Each Jurisdiction

	2015							2017						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
States														
Arizona	🟨	🟨	🟨	🟨	▒	▒	🟨	🟨	🟨	🟨	🟨	▒	▒	🟨
Arkansas	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
California	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Colorado	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟨	▒	▒	▒	▒	▒	🟨
Connecticut	🟨	🟨	🟨	🟨	🟨	▒	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛
Delaware	🟨	🟨	▒	▒	🟨	🟨	🟨	▒	▒	▒	▒	▒	🟨	🟨
Florida	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Hawaii	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Illinois	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Iowa	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Kentucky	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Maine	🟨	▒	▒	▒	▒	🟨	🟨	🟨	▒	▒	▒	▒	🟨	🟨
Maryland	🟨	🟨	🟨	🟨	🟨	🟨	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛
Michigan	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Nebraska	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Nevada	🟨	🟨	🟨	🟨	🟨	▒	🟨	🟨	▒	▒	▒	🟨	▒	🟨
New Hampshire	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟨	🟨	▒	▒	🟨	▒	🟨
New York	🟨	▒	▒	▒	🟨	🟨	🟨	🟨	▒	▒	▒	🟨	🟨	🟨
North Carolina	🟨	🟨	🟨	🟨	▒	🟨	🟨	🟨	🟨	🟨	🟨	▒	🟨	🟨
North Dakota	🟨	🟨	🟨	🟨	🟨	▒	🟨	🟨	🟨	🟨	🟨	🟨	▒	🟨
Oklahoma	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Pennsylvania	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Rhode Island	🟨	🟨	🟨	🟨	🟨	▒	🟨	🟨	🟨	🟨	🟨	🟨	▒	🟨
South Carolina	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟨	🟨	🟨	🟨	🟨	🟨	🟨
West Virginia	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Wisconsin	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Wyoming	🟨	🟨	🟨	🟨	🟨	🟨	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛
Districts														
Bronx, NY	🟨	▒	▒	▒	🟨	🟨	🟨	🟨	▒	▒	▒	🟨	🟨	🟨
Brooklyn, NY	🟨	▒	▒	▒	🟨	🟨	🟨	🟨	▒	▒	▒	🟨	🟨	🟨
Broward County, FL	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Chicago, IL	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Duval County, FL	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Fort Worth, TX	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Manhattan, NY	🟨	▒	▒	▒	🟨	🟨	🟨	🟨	▒	▒	▒	🟨	🟨	🟨
Miami–Dade County, FL	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Orange County, FL	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Queens, NY	🟨	▒	▒	▒	🟨	🟨	🟨	🟨	▒	▒	▒	🟨	🟨	🟨
San Diego, CA	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Staten Island, NY	🟨	▒	▒	▒	🟨	🟨	🟨	🟨	▒	▒	▒	🟨	🟨	🟨

Note. Item 1: fruit; Item 2: salad; Item 3: potatoes; Item 4: carrots; Item 5: dentist; Item 6: asthma; Item 7: height. Items excluded from surveys administered in each jurisdiction are indicated with gray boxes, while included items are indicated with yellow boxes. Jurisdictions that did not participate in the survey year are indicated by black boxes. Boldfaced states and districts contained all 7 screener items in the 2017 data, and were used for all main analyses; analyses using all 2017 data (regardless of screener items included) are in Appendix B.

Following the boosted regression, each student is ranked by how likely they are to be a mischievous responder based on their response combination to the screener items. We use weighted linear probability models to obtain estimates of LGBQ-heterosexual disparities, with ordinal values recoded as continuous and include location fixed effects, using the

TABLE 2

Demographic Characteristics of the Weighted Sample, by Reported Sex and Screener Threshold, Pooled Youth Risk Behavior Survey (YRBS) 2015, Pooled YRBS 2017, and the Subset of YRBS 2017 Who Were Administered All Seven Screener Items

	2015		2017		Full-Screener Subsample	
	Males (*N* = 72,641)	Females (*N* = 76,319)	Males (*N* = 52,753)	Females (*N* = 55,340)	Males (*N* = 25,036)	Females (*N* = 26,488)
Sexual identity						
Heterosexual	91.77	82.82	89.99	79.57	91.02	80.66
Gay or lesbian	2.49	2.07	2.95	2.64	2.90	2.48
Bisexual	2.88	9.94	3.20	11.99	2.93	11.72
Not sure	2.86	5.17	3.86	5.80	3.16	5.14
Race						
White	46.58	45.12	47.69	48.27	47.50	48.58
Black or African American	14.62	15.55	15.08	15.22	14.66	14.99
Hispanic/Latino	27.14	28.32	26.49	26.35	27.14	26.38
All other races	11.66	11.02	10.74	10.16	10.69	10.05
Grade						
9th grade	27.50	26.84	26.61	26.09	26.28	25.98
10th grade	25.75	25.62	26.13	26.10	26.11	26.02
11th grade	23.96	24.04	24.15	24.45	24.26	24.56
12th grade	22.78	23.50	23.11	23.23	23.35	23.44
Age						
12 years old or younger	0.29	0.32	0.37	0.30	0.34	0.28
13 years old	0.47	0.47	0.40	0.36	0.17	0.12
14 years old	12.29	13.49	11.67	12.84	10.68	12.07
15 years old	25.50	26.11	24.73	25.36	24.47	25.31
16 years old	24.94	24.88	25.89	25.93	26.35	25.98
17 years old	23.14	22.80	23.05	23.40	23.39	23.55
18 years old or older	13.37	11.92	13.89	11.80	14.59	12.68

full analytic data set. Then, we remove the top 1% of students (based on likely mischievousness) and reestimate the disparities. We repeat this process, sequentially removing the next 1% of data and reestimating the disparities, until 25% of the data have been removed. Additional details on the methods for Study 1 (and Studies 2 and 3) can be found in Appendix A.

Study 2: Direct Replication of Cimpian et al.'s (2018) Analysis of the Relationship Between Item Response-Option Extremity and Screening Effects

We directly replicate Cimpian et al.'s (2018) analysis exploring if the variation in screening effects across the 20 outcomes is related to how frequently respondents select the most extreme response options. Mischievous responders often choose extreme response options (Fan et al., 2006; Furlong et al., 2004; Furlong et al., 2017; Robinson-Cimpian, 2014), and items with fewer respondents overall selecting these options are then more susceptible to bias. We use a random effects model to predict the reduction in the estimate of LGBQ-heterosexual disparities between the model using all data and a given model with 1% to 25% of potential mischievous responders removed.

Study 3: Comparison of Post Hoc Mischievousness Reduction Techniques

While Study 1 detects mischievousness using *one* method—the most computationally complex and recently applied method—and Study 2 builds off of that detection method, Study 3 compares *four* methods for detecting mischievousness. Here, we briefly describe the various methods we will compare, then we discuss how we compare them; additional details are in Appendix A.

Method 1: Boosted Regression. This method is described above in Study 1.

General Notes for Methods 2 to 4. In all of the following approaches (i.e., anything other than the boosted regressions), the researchers must prespecify which response-options are considered tempting to mischievous responders. Note that this requires different assumptions than the boosted

TABLE 3

Demographic Characteristics of the Weighted Sample, by Reported Sex and Screener Threshold, Subset of Youth Risk Behavior Survey 2017 Who Were Administered all Seven Screener Items

	Males (Unweighted N = 25,036)				Females (Unweighted N = 26,488)			
	Full Sample	5% Screened Out	10% Screened Out	25% Screened Out	Full Sample	5% Screened Out	10% Screened Out	25% Screened Out
Sexual identity								
Heterosexual	91.02	91.93	92.56	93.37	80.66	81.36	81.99	83.53
Gay or lesbian	2.90	2.62	2.39	2.21	2.48	2.35	2.24	1.98
Bisexual	2.93	2.70	2.56	2.37	11.72	11.38	11.05	10.42
Not sure	3.16	2.74	2.50	2.05	5.14	4.90	4.73	4.07
Race								
White	47.50	47.53	47.55	47.22	48.58	48.81	48.99	50.09
Black or African American	14.66	14.45	14.15	13.59	14.99	14.44	14.03	12.59
Hispanic/Latino	27.14	27.31	27.61	28.23	26.38	26.60	26.77	27.04
All other races	10.69	10.71	10.69	10.97	10.05	10.14	10.22	10.28
Grade								
9th grade	26.28	26.17	26.02	25.62	25.98	25.79	25.63	25.34
10th grade	26.11	26.25	26.26	26.40	26.02	26.08	26.24	26.63
11th grade	24.26	24.36	24.39	24.50	24.56	24.70	24.77	24.86
12th grade	23.35	23.22	23.33	23.48	23.44	23.43	23.36	23.17
Age								
12 years old or younger	0.34	0.25	0.22	0.21	0.28	0.24	0.21	0.17
13 years old	0.17	0.17	0.16	0.13	0.12	0.11	0.11	0.10
14 years old	10.68	10.71	10.68	10.63	12.07	12.10	12.13	11.91
15 years old	24.47	24.53	24.57	24.59	25.31	25.31	25.32	25.83
16 years old	26.35	26.45	26.62	26.69	25.98	26.01	26.14	26.52
17 years old	23.39	23.51	23.45	23.52	23.55	23.64	23.62	23.26
18 years old or older	14.59	14.38	14.31	14.23	12.68	12.59	12.46	12.20

Note. Our sensitivity analysis approach was to remove a fixed number (i.e., unweighted amount) of observations at each removal step (e.g., 5% screened out, 10% screened out). The percentages in the table reflect the weighted sample that remains at each of the selected screening thresholds.

regression, which is important in two ways. First, while the boosted regression requires prespecifying *items* to consider (e.g., how often do you eat carrots?), the following methods require prespecifying *item response-options* to weight (e.g., eating carrots "4 or more times a day"). Second, the boosted regression will ignore the prespecified items if they are not helpful in differentiating between likely mischievous and nonmischievous respondents and it will give more weight to the items that are more helpful in differentiating; this is because there is no prespecification of how much to weight these items. By contrast, the following methods implicitly preweight the contributions of each pre-specified item response-option. That is, even if eating carrots "4 or more times a day" does not distinguish between reporting to be in the minority versus majority group, it will contribute to the ranking of likely mischievousness simply because the response option is a low-frequency choice and was preidentified by the researchers as an unusual response option. Thus, assumptions about which items and which specific item response-options are selected play a larger role in Methods 2 through 4.

Regarding our prespecification of tempting item response-options for this analysis, the Cimpian et al. (2018) analyses suggest a set of response-options to the screener items that are unusual and suggestive of likely mischievous responders. Based on that study, Table 4 presents the list of response-options we deem suggestive of mischievous responding in the 2017 YRBS.[2] (These item response-options in Table 4 are also used for Methods 3 and 4.)

Method 2: Unconditional Probability-Based Ranking. This method multiplies all of the unconditional probabilities of the prespecified screener item responses-options together and then ranks observations by their multiplied probabilities (Robinson-Cimpian, 2014). Individuals who selected most of the lowest probability researcher-specified mischievous response-options are sequentially removed from the data at 1% intervals and disparities are reestimated.

TABLE 4

Prespecified Screener Items and Extreme Response-Options in the Subset of Youth Risk Behavior Survey 2017 Who Were Administered All Seven Screener Items

Item (any bolding appears in the actual survey instrument)	Prespecified tempting/unusual response-option	Frequency which the response-option is selected
How tall are you without your shoes on?	The Centers for Disease Control and Prevention recodes “biologically implausible” values of height to missing; therefore, we will consider missing values as unusual (consistent with evidence from Cimpian et al., 2018)	8.11%
During the past 7 days, how many times did you eat **fruit**? (Do **not** count fruit juice.)	4 or more times per day	5.27%
During the past 7 days, how many times did you eat **green salad**?	4 or more times per day	1.33%
During the past 7 days, how many times did you eat **potatoes**? (Do **not** count french fries, fried potatoes, or potato chips.)	4 or more times per day	1.29%
During the past 7 days, how many times did you eat **carrots**?	4 or more times per day	1.05%
When was the last time you saw a dentist for a check-up, exam, teeth cleaning, or other dental work?	Never	2.23%
Has a doctor or nurse ever told you that you have asthma?	Not sure	5.16%

Note. The wording (and bolding) in the first two columns is taken directly from the 2017 Youth Risk Behavior Survey.

Method 3: Count Based. The count-based removal of suspected mischievous responders is similar to the probability-based approach, except each prespecified low-frequency response-option receives equal weight. More simply, this approach just tallies up the number of low-frequency responses provided to the items in Table 4. The more low-frequency response options a respondent provides, the more likely they are to be mischievous, and these individuals are removed sequentially in 1% intervals as in all the methods described above.

Method 4: Regression Adjustment. The regression adjustment approach uses the values of P from the probability-based approach (see Appendix A), but statistically conditions on a function of P rather than remove observations based on the ranking of P. That is, there is no data removal and no reweighting of observations in any way in this method, which separates it from all the methods described above; there is simply a regression-based covariate adjustment. We explore the consequences of different functional forms of P on how the estimates of the disparities: linear P, natural log of P (used in Fish & Russell, 2018), and the quartic of the natural log of P (used in Robinson-Cimpian, 2014).

Comparing the Methods. First, for each method (except the regression-based ones [Method 4]), we estimate the change in the average LGBQ-heterosexual disparities from the model using the full analytic data set to the model that screens out the top 1% of data. This average change is the precision-weighted average of the change in the 20 outcomes and adjusted for the covariance matrix in the changes (e.g., not treating changes in suicidal-ideation disparities as independent from changes in suicide-planning disparities). At each percentage of data removal, we test whether the removal of suspect data had a larger effect via one data-validity method relative to the others. We are able to see when in the sensitivity analysis (i.e., data removal process) the various methods yield the same results. To compare Method 4 (i.e., the regression-based adjustments), which each yield only one estimate (as opposed to the range of estimates produced by Methods 1–3), we show where the various regression-based adjustments fall in relation to the various other methods.

Hypotheses

Hypotheses for Study 1 (Preregistered Replication)

Based primarily on the recent work of Cimpian et al. (2018), we make several predictions. We predict the removal of potentially mischievous responders, as identified through the boosted regression, will lead to significant reductions in LGBQ-heterosexual disparities averaged over the 20 outcomes. We predict these reductions will be significant as soon as the top 1% of observations are removed. We predict the reductions will be larger among males than among females, who have been found to demonstrate less mischievousness in surveys (Cimpian et al., 2018; Fan et al., 2006). We do not make predictions about the individual outcomes, but only about the average of the 20 outcomes.

Hypotheses for Study 2 (Preregistered Replication)

Based on the findings of Cimpian et al. (2018), we expect to replicate their findings that item response-option extremity is predictive of the magnitude of the disparity reductions experienced when screening out potentially mischievous responder, for both males and females.

Hypotheses for Study 3 (Exploratory)

We make no a priori predictions for Study 3. Based on prior research relating boosted regressions to propensity score matching (McCaffrey, Ridgeway, & Morral, 2004) and on the efficiency of boosted regressions over previous supervised machine-learning methods (Hastie, Tibshirani, & Friedman, 2017), we do suspect the boosted regression to be most efficient in identifying potentially mischievous responders (in this case, if mischievous responders were biasing estimates upward and if the methods to detect them were not introducing bias themselves, efficiency would translate into smaller disparity estimates reached when removing fewer observations). However, we refrain from making strong predictions, and we view this study as exploratory.

Results

Study 1

As hypothesized, the results of Study 1 indicate that the removal of potentially mischievous responders leads to a significant reduction in average estimated disparities. This reduction is significant when just 1% of observations are removed, and the reduction is much larger for males than females. We focus here on results for the subset of jurisdictions that administered all seven screener items. Analyses of the full data set demonstrate similar trends and are available in Appendix B (Figures B17–B19).

When using all of the data, the average LGBQ-heterosexual youth health outcome disparity was 0.33 standard deviations (*SD*s) (95% confidence interval [CI] [0.25, 0.41]) among males (see Table 5). Here, we focus on the results of the boosted regression, but Table 5 shows results for all methods. When we removed the top 1% of observations identified by the boosted regression as providing the most unusual response patterns to the screener items, the estimated disparity reduced to 0.30 *SD*s (95% CI [0.22, 0.38]), and the change in the disparity was itself statistically significant (as indeed, all the changes are that are presented in Table 5). The estimate decreased to 0.25 *SD*s (95% CI [0.17, 0.33]) when removing the top 5%, decreased to 0.21 *SD*s (95% CI [0.13, 0.30]) when removing 10%, and decreased to 0.16 *SD*s (95% CI [0.08, 0.24]) when removing 25%. That is, the average of the male LGBQ-heterosexual disparities was *cut in half* when removing the top 25% of students ranked by likely mischievousness, yet neither did this data removal substantially alter any demographics of the data set (suggesting good generalizability; see Table 3) nor did it appreciably reduce precision of the estimated disparities.

Among females, the average LGBQ-heterosexual estimated outcome disparity was 0.25 *SD*s (95% CI [0.18, 0.33]). Similar to the 2015 results presented in Cimpian et al. (2018), the changes in estimated disparities are much smaller among females than males. When removing the top 25% of observations, the estimated disparities decreased to 0.21 *SD*s (95% CI [0.12, 0.30]).

Both males and females demonstrated very similar patterns to the respondents of the 2015 YRBS. For both groups, estimates using the full sample were somewhat larger in 2015 than in 2017, with the average LGBQ-heterosexual disparity at 0.37 *SD*s (95% CI [0.29, 0.45]) for males and 0.31 *SD*s (95% CI [0.23, 0.38]) for females (Cimpian et al., 2018). Disparity estimates for males dropped much more substantially than for females in both survey administrations. In both 2015 and 2017, the difference for males between the full sample estimate and that based on 25% screened out was 0.17 (95% CI [0.11, 0.23] in 2015, and 95% CI [0.10, 0.24] in 2017). For females, differences were relatively smaller than males in both years, with the difference between the full estimate and that of the 25% screened out 0.07 (95% CI [0.03, 0.11]) in 2015 and 0.04 (95% CI [0.01, 0.08]) in 2017. While the estimated disparities are of course not identical, the patterns presented here using the 2017 YRBS are consistent with those Cimpian et al. (2018) identified in the 2015 YRBS.

Study 2

The results of Study 2 are also consistent with the findings of Cimpian et al. (2018). There is considerable variability in how screening affects estimated disparities across the 20 outcomes, and items with extreme response-options are predictive of the magnitude of disparity reductions.

As illustrated in Figures 1 and 2, while estimates of disparities for outcomes related to bullying and suicidal ideation were generally relatively stable for both boys and girls, disparities for drug- and alcohol-related outcomes were affected by the removal of potentially mischievous responders much more dramatically, particularly for boys. For example, among boys, the LGBQ-heterosexual boosted regression-based estimated disparity for heroin use showed an immediate steep decline, dropping from 0.55 *SD*s to 0.07 *SD*s on removal of the top 25% of responders. Disparities in alcohol and ecstasy

TABLE 5

Estimates and 95% Confidence Intervals for Average LGBQ-Heterosexual Youth Health Disparities Across 20 Outcomes, Subset of Youth Risk Behavior Survey 2017 Who Were Administered All Seven Screener Items

	Males		Females	
	Average LGBQ-Heterosexual disparity	Change in the average disparity from model using the full sample to a model using a screened sample or regression adjustment	Average LGBQ-Heterosexual disparity	Change in the average disparity from model using the full sample to a model using a screened sample or regression adjustment
Full sample	0.33 [0.25, 0.41]	—	0.25 [0.18, 0.33]	—
Boosted regression-based removal of data				
1% screened	0.30 [0.22, 0.38]	0.02 [0.01, 0.04]	0.24 [0.16, 0.32]	0.01 [0.01, 0.02]
5% screened	0.25 [0.17, 0.33]	0.08 [0.05, 0.11]	0.22 [0.14, 0.30]	0.03 [0.01, 0.05]
10% screened	0.21 [0.13, 0.30]	0.12 [0.07, 0.16]	0.22 [0.13, 0.30]	0.04 [0.01, 0.06]
15% screened	0.19 [0.11, 0.27]	0.14 [0.08, 0.19]	0.21 [0.13, 0.30]	0.04 [0.01, 0.07]
20% screened	0.17 [0.09, 0.25]	0.16 [0.10, 0.22]	0.21 [0.12, 0.30]	0.04 [0.01, 0.07]
25% screened	0.16 [0.08, 0.24]	0.17 [0.10, 0.24]	0.21 [0.12, 0.30]	0.04 [0.01, 0.08]
Probability-based removal of data				
1% screened	0.31 [0.23, 0.39]	0.02 [0.01, 0.02]	0.25 [0.17, 0.33]	0.01 [0.00, 0.01]
5% screened	0.29 [0.21, 0.37]	0.04 [0.03, 0.05]	0.23 [0.15, 0.31]	0.03 [0.02, 0.03]
10% screened	0.27 [0.19, 0.35]	0.06 [0.05, 0.08]	0.23 [0.15, 0.31]	0.03 [0.01, 0.04]
15% screened	0.26 [0.18, 0.34]	0.07 [0.05, 0.09]	0.22 [0.14, 0.31]	0.03 [0.02, 0.04]
20% screened	0.26 [0.18, 0.34]	0.07 [0.05, 0.09]	0.22 [0.13, 0.30]	0.04 [0.02, 0.05]
25% screened	0.23 [0.15, 0.31]	0.10 [0.08, 0.13]	0.22 [0.13, 0.30]	0.04 [0.02, 0.05]
Count-based removal of data				
1% screened	0.31 [0.23, 0.39]	0.02 [0.01, 0.02]	0.25 [0.17, 0.33]	0.01 [0.00, 0.01]
5% screened	0.29 [0.21, 0.37]	0.04 [0.03, 0.05]	0.24 [0.16, 0.32]	0.01 [0.01, 0.02]
10% screened	0.29 [0.21, 0.37]	0.04 [0.03, 0.05]	0.24 [0.16, 0.32]	0.01 [0.01, 0.02]
15% screened	0.29 [0.21, 0.37]	0.04 [0.03, 0.05]	0.24 [0.16, 0.32]	0.01 [0.01, 0.02]
20% screened	0.29 [0.21, 0.37]	0.04 [0.02, 0.06]	0.22 [0.13, 0.30]	0.04 [0.02, 0.05]
25% screened	0.23 [0.15, 0.31]	0.10 [0.08, 0.13]	0.22 [0.13, 0.30]	0.04 [0.02, 0.05]
Regression adjustment				
Linear	0.31 [0.23, 0.40]	0.01 [0.01, 0.02]	0.25 [0.17, 0.33]	0.01 [0.00, 0.01]
Nonlinear	0.31 [0.23, 0.39]	0.02 [0.02, 0.03]	0.24 [0.17, 0.32]	0.01 [0.01, 0.01]
Quartic	0.31 [0.23, 0.39]	0.02 [0.02, 0.03]	0.24 [0.17, 0.32]	0.01 [0.01, 0.01]

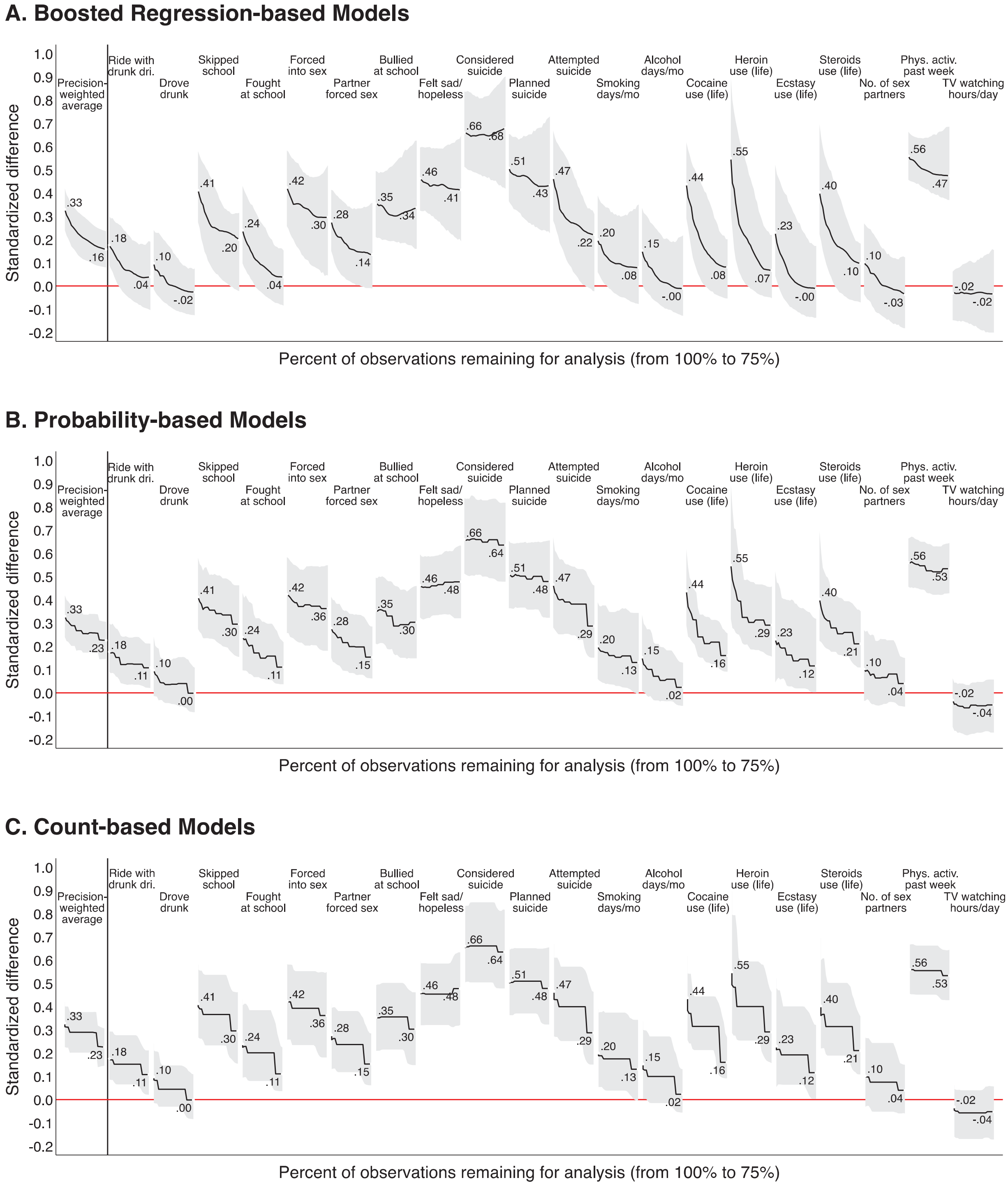
Note. LGBQ = lesbian, gay, bisexual, or questioning. All estimates are reported as standardized differences and can be interpreted using typical effect size standards. All estimates (both overall and difference) are statistically significant at $p < .004$. This table presents a concise subset of screening values. Male $N = 25{,}036$, female $N = 26{,}488$.

use also dropped to 0.00 *SD*s among boys when removing the top 25% of responders. In contrast, the estimated disparity for reporting being bullied at school is virtually unchanged, moving only from 0.35 *SD*s to 0.34 *SD*s. Thus, it is unlikely that disparities in bullying and other relatively stable outcomes are driven by mischievous responders, whereas drug-related outcomes in particular are susceptible to their influence. While LGBQ-heterosexual disparity estimates among girls were generally more stable than the boys, similar patterns in drug- and alcohol-related outcomes are evident.

In Figure 3, we explore the relationship between item response-option extremity and the average change in the estimated LGBQ-heterosexual disparity based on the boosted regression approach (see Table 6 for the extreme response options for each outcome). For both boys and girls, the smaller the proportion of respondents choosing the most extreme response option, the larger the change in the disparity (*p*s < .001). In both cases, drug-related outcomes notably have both very small numbers of students endorsing the most extreme response options and also demonstrate large changes in estimated disparities. In contrast, outcomes with relatively more commonly selected extreme response options also tend to have smaller changes in estimated disparities (e.g., bullied at school, physical activity). As with Study 1, our results for Study 2 demonstrate similar patterns to those found by Cimpian et al. (2018) with the 2015 YRBS.

Study 3

In our exploratory analyses, we compared the average disparities estimated via the different methods for addressing potentially mischievous responders. The results differ



A. Boosted Regression-based Models

B. Probability-based Models

C. Count-based Models

FIGURE 1. *Average LGBQ-Heterosexual disparity among reported males in the full-screener subsample, by model, outcome, and percent of observations screened out.*
Note. LGBQ = lesbian, gay, bisexual, or questioning. The shaded areas represent asymmetrical 95% confidence *intervals* (CIs), constructed via 1999 bootstrapped samples. If the shaded area does not cross the horizontal red line at zero, the disparity is statistically significant ($p < .05$).

somewhat by gender: Among males, there are clear differences among the methods; but among females, where likely mischievousness was tempered, so were the differences between the methods. In all cases, the boosted regression approach to identification followed by data removal led to statistically significantly smaller disparity estimates than did any of the regression-based adjustment methods (i.e., the methods that did not remove any data). Moreover, among

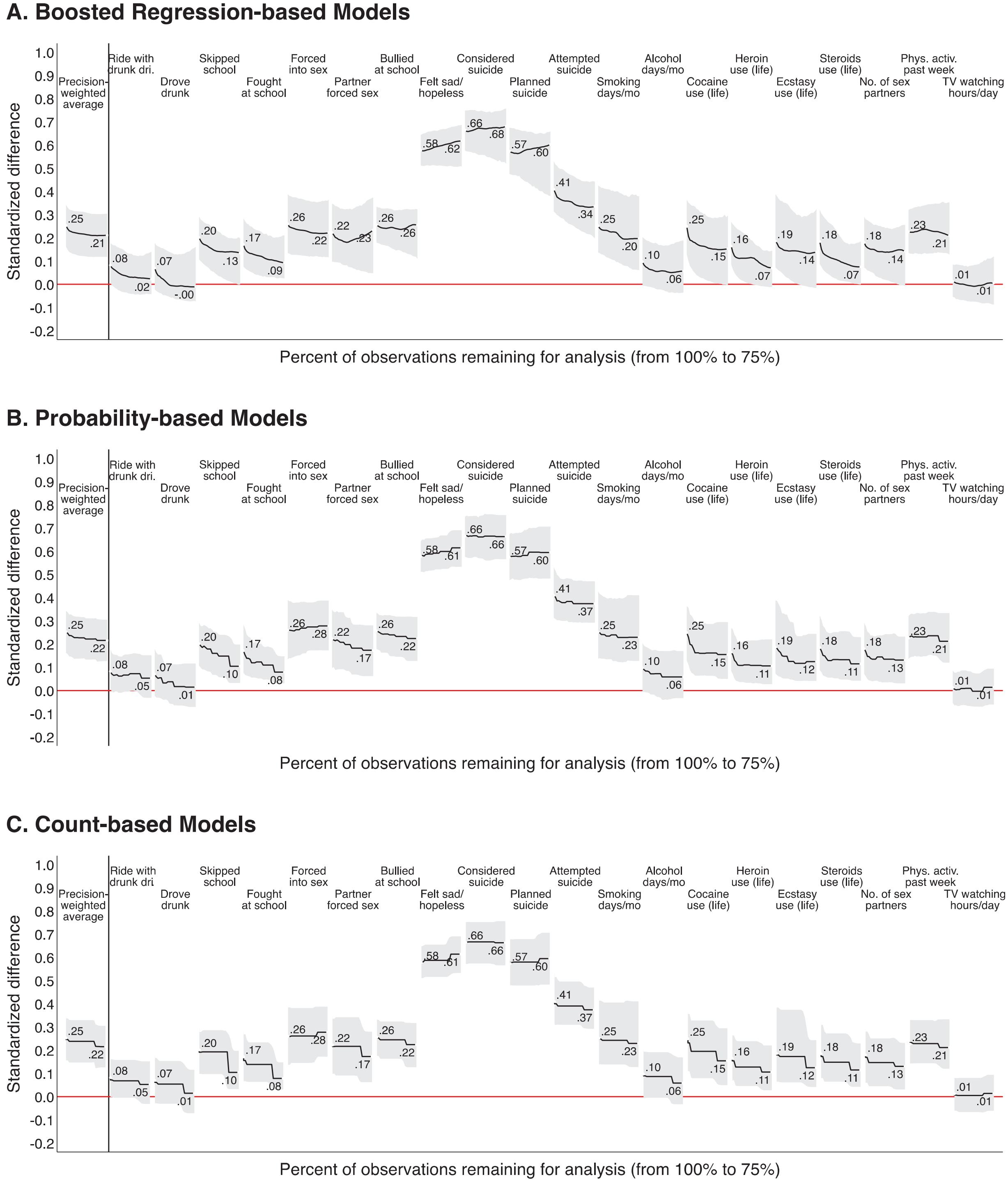


FIGURE 2. *Average LGBQ-Heterosexual disparity among reported females in the full-screener subsample, by model, outcome, and percentage of observations screened out.*

Note. LGBQ = lesbian, gay, bisexual, or questioning. The shaded areas represent asymmetrical 95% confidence intervals (CIs), constructed via 1999 bootstrapped samples. If the shaded area does not cross the horizontal red line at zero, the disparity is statistically significant ($p < .05$).

males, the boosted regression approach followed by data removal led to smaller disparity estimates than did any of the other mischievous responder detection techniques followed by data removal. Figure 4 is a graphical illustration of the information presented in Table 5. In Figure 4, we see that the disparity line from the boosted regression is below every

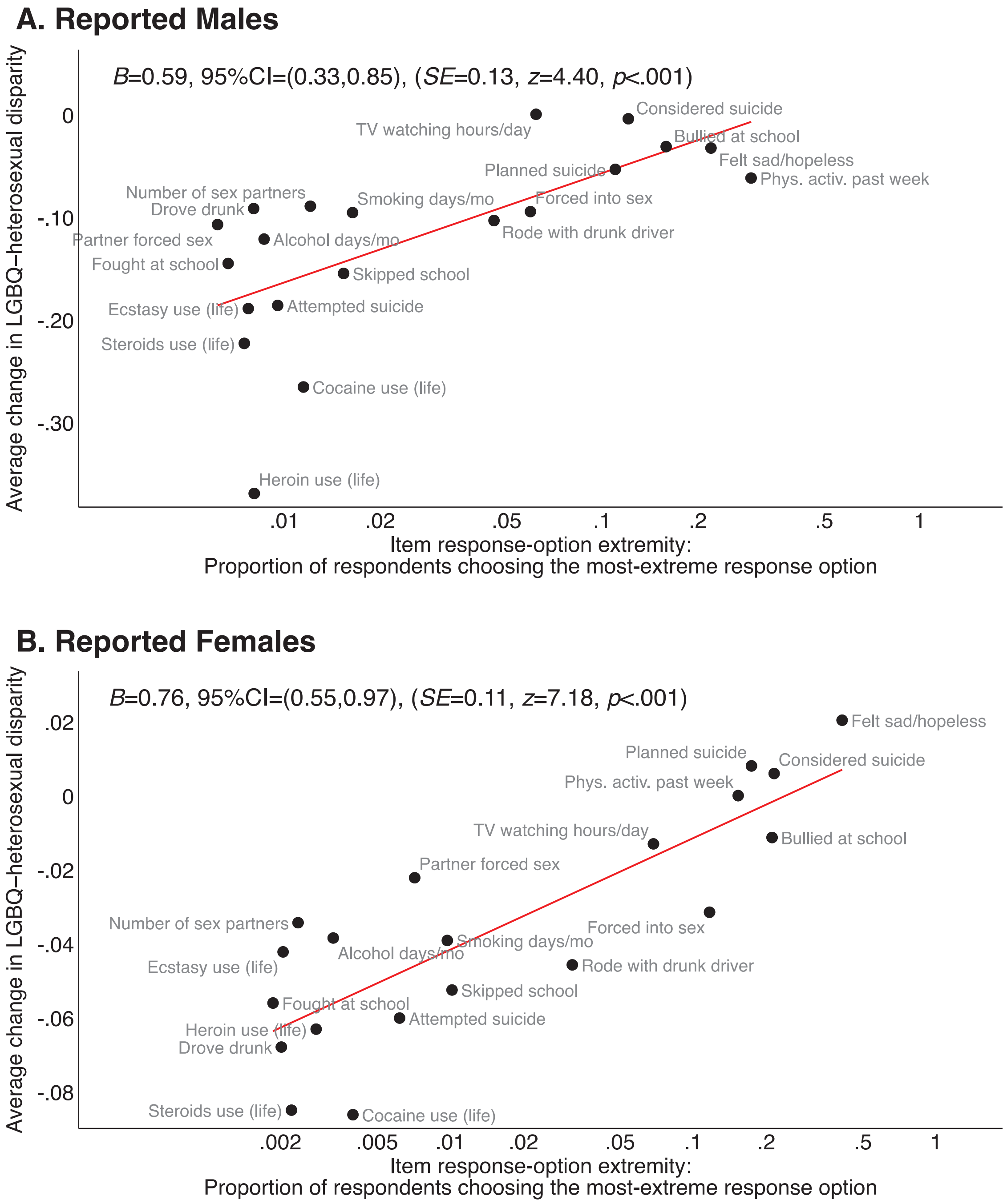


FIGURE 3. *Relationship between how much an item-level disparity changed when screening mischievous responders and the item response-option extremity, in the full-screener subsample.*

Note. More extreme item response-options (to the left on the *x*-axis) mean fewer respondents chose the most extreme option (e.g., using heroin "40 or more times"), which corresponds to larger average changes in the disparities when screening out mischievous responders.

other data removal method, showing that it leads to the smallest average disparities at every estimation point.

Figure 5 directly compares the boosted regression estimates to each of the other estimates: Each line in Figure 5 represents a "difference in differences" of sorts, where the LGBQ-heterosexual difference from one method (e.g., probability-based approach, linear regression approach) is subtracted from the LGBQ-heterosexual difference from the boosted regression. The colored areas are 95% CIs for those difference-in-differences estimates, estimated via 1999 bootstrapped samples. Among males, the estimates from the boosted regression-based approach are statistically

TABLE 6

Outcome Items and Extreme Response-Options in the Subset of Youth Risk Behavior Survey 2017 Who Were Administered All Seven Screener Items

		Males		Females	
Outcome	Extreme Response-Option	Full Sample (%)	Full-Screener Subsample (%)	Full Sample (%)	Full-Screener Subsample (%)
Rode with drunk driver	6 or more times in past 30 days	4.11	4.56	2.81	3.18
Drinking and driving	6 or more times in past 30 days	0.89	0.80	0.27	0.20
Safety concerns at school	6 or more days in past 30 days	1.39	1.53	0.95	1.02
Physical fighting at school	12 or more times in past 12 months	0.71	0.66	0.19	0.19
Forced sexual intercourse	Yes (ever)	5.13	5.94	10.99	11.67
Sexual dating violence	6 or more times in past 12 months	0.88	0.62	0.86	0.71
Bullying at school	Yes (past 12 months)	15.88	15.90	21.39	21.11
Feeling sad or hopeless	Yes, almost every day for 2 weeks or more in a row (past 12 months)	21.56	21.99	39.96	41.02
Considered suicide	Yes (past 12 months)	12.00	12.08	20.81	21.53
Made a suicide plan	Yes (past 12 months)	10.73	10.99	17.04	17.37
Attempted suicide	6 or more times in past 12 months	0.99	0.95	0.58	0.62
Current cigarette use	All 30 days (past 30 days)	1.69	1.64	0.95	0.97
Current alcohol use	All 30 days (past 30 days)	0.96	0.86	0.31	0.33
Cocaine use (ever)	40 or more times	1.06	1.15	0.37	0.40
Heroin use (ever)	40 or more times	0.89	0.80	0.27	0.28
Ecstasy use (ever)	40 or more times	0.87	0.77	0.22	0.20
Steroid use (ever)	40 or more times	0.73	0.75	0.21	0.22
Current sexual activity	6 or more people	1.25	1.21	0.24	0.24
Physical activity	7 days (past 7 days)	29.01	29.41	15.71	15.32
Television watching	5 or more hours per day	6.28	6.19	6.63	6.86

significantly different from those of all other approaches. For example, if 25% of the data were removed following the boosted regression identification method, the male LGBQ-heterosexual average disparity would be 0.16 *SD*s smaller than the estimate based on a regression-adjustment approach. Not only is that difference between methods statistically significant it also represents the practical difference of almost *half* of the unadjusted average LGBQ-heterosexual disparity; that is, the original male LGBQ-heterosexual average disparity using all data was 0.33 *SD*s, which reduced slightly to 0.31 *SD*s using the regression adjustment, but was cut in more than half to 0.16 *SD*s in the final boosted regression estimate.

Taken together, the results of Study 3 suggest that, if these methods are indeed identifying mischievous responders who are biasing disparity estimates, then (1) data removal eliminates more of the bias than do covariate adjustment approaches and (2) among the data removal approaches, the boosted regression approach to identifying likely mischievous responders leads to faster bias removal than do either the probability- or count-based approaches, reducing bias while being able to retain the greatest amount of observations. The differences between the approaches is more consequential when there is more potential for bias, such as in the case of males in the YRBS.

Discussion

This article replicates Cimpian et al. (2018) and finds consistent results regarding how potentially mischievous responders affect LGBQ-heterosexual disparities in Study 1. Furthermore, it adds a new empirical test for the theory of which items and response-options are most likely affected by mischievous responders in Study 2, helping survey developers plan in advance to mitigate the effects of mischievous responders, as well as helping applied researchers identify mechanisms linking these patterns together. For example, it may be at first confusing why suicidal ideation is not affected by mischievous responders but suicide attempts are, but this mechanism of item response-option extremity helps make

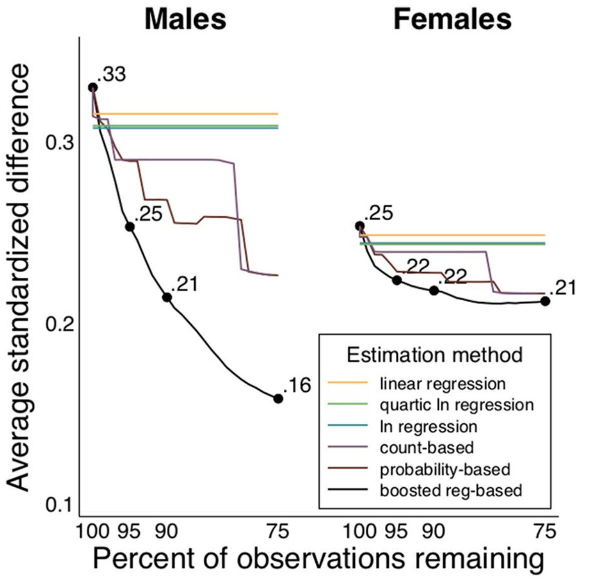


FIGURE 4. *Average standardized LGBQ-Heterosexual disparities by gender, estimation method, and percentage of observations remaining, in the full-screener subsample.*
Note. LGBQ = lesbian, gay, bisexual, or questioning. The linear regression, natural log regression, and quartic natural log regression do not remove any observations; hence, they are represented by horizontal lines. All estimates are significantly different from zero, therefore, we do not present confidence intervals.

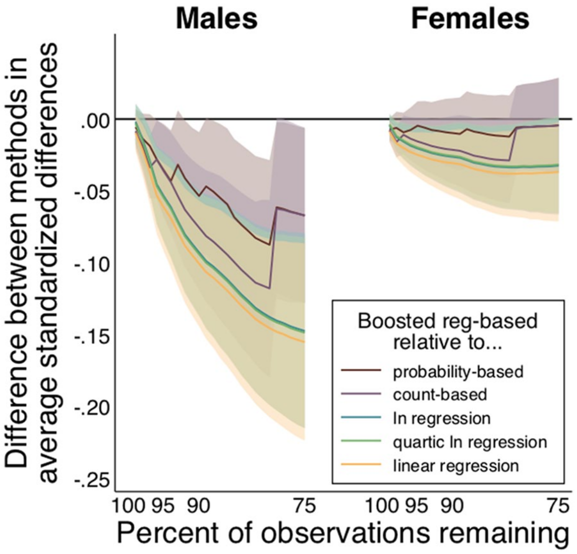


FIGURE 5. *Difference in average standardized LGBQ-Heterosexual disparities from the boosted regression approach, by gender, estimation method (compared with the boosted regression), and percentage of observations remaining, in the full-screener subsample.*
Note. LGBQ = lesbian, gay, bisexual, or questioning. Each line represents the difference between the estimated disparity in the boosted regression to an estimated disparity from a different estimation method (e.g., probability-based identification followed by data removal). The colored shaded areas corresponding to the colored lines represent 95% confidence intervals (CIs). If the colored 95% CI does not the zero line, then the boosted regression estimate yielded a statistically significantly smaller LGBQ-heterosexual disparity estimate. For those mapping this figure onto Figure 4, this figure is just *subtracting* the line from a given estimation method from the line of the boosted regression.

sense of this pattern and an even broader set of patterns. That is, more students overall reported thinking about suicide than attempting it (especially repeated attempts); we replicate Cimpian et al.'s (2018) finding that likely mischievous responders will have undue influence on items where relatively fewer students choose the extreme response (e.g., suicide attempts) and that their influence is diluted when more students choose the extreme response (e.g., suicidal ideation). In addition to the preregistered replications, this article adds comparisons across the methods for identifying and either removing or adjusting for potentially mischievous responders in Study 3. The differences between the methods deserve additional focus, with growing concerns about ensuring data validity. We begin by discussing practical implications for nonresearchers, and then discuss the broader issue of research transparency.

Practical Implications for Interpreting Results

If researchers follow our suggestions, then practitioners, policy makers, and education decision makers would encounter articles and reports with a range of estimates for each outcome. This can be daunting and confusing, especially if the results from one model contradict those of another model. Importantly, if the results are inconsistent across the models, then the practitioners/policy makers/decision makers should use extreme care when interpreting the research studies and making real-world decisions (and researchers should clarify any data-validity concerns). We illustrate this point with a couple of examples from the current article. First, the disparity estimates were more stable across the models among females than among males, suggesting to decision makers that data validity may be less of an issue when reviewing survey data on females, and correspondingly, their data-based decisions regarding females are less sensitive to the specific model choices. The results using data on males, however, were more sensitive to modeling assumptions, thus decision makers will need to especially weight the plausibility of these assumptions when deciding how to proceed with policy and practice for males. Second, even among males, though, some disparity estimates were more stable than others. For instance, LGBQ males reported about one third of an *SD* higher likelihood of being bullied no matter the modeling assumptions; by contrast, the significance and/or magnitude of the estimated disparity depends on modeling assumptions for outcomes like fighting at school, skipping school, and a wide range of alcohol and drug uses. These patterns of (in)stability across the estimates may lead decision makers to conclude that the bullying disparity is not substantially influenced by potentially invalid data, and therefore, may require action on the part of educators to reduce this disparity. The data are less clear on the other outcomes, but we would not know that if we did not perform these sensitivity analyses. That is, these other outcomes may also require action, but the results of the survey data are inconclusive as to whether a disparity exists or its magnitude because the estimates change so much from model to model. Practically speaking, it is important to know if these estimated disparities are sensitive to modeling assumptions before resources are devoted to addressing and monitoring these outcomes by group.

Data Validity and Research Transparency

In the vein of the theoretical critiques in the recent special issue of *Educational Researcher* (see, e.g., Brockenbrough, 2017; Cimpian, 2017; Love, 2017; Mayo, 2017), this article also challenges—empirically—the assumptions implicit in much quantitative education research on LGBQ youth (see also, Robinson-Cimpian, 2014). Yet, this work pushes the empirical work a step further by providing direct comparisons of methods used for assessing data validity, and in doing so, illustrates that even seemingly similar work intended to reduce bias can lead to different conclusions based on the assumptions the researchers make throughout the analysis stage. This sort of methodological questioning extends well beyond the LGBQ (and LGBTQ+) research literature, to disparities related to other majority-minority comparisons, and even to the broader discussion of general data validity. In each study, researchers are making choices about overall and differential data validity.

We can think of these various researcher choices—perhaps charitably—as confronting and reducing messiness for a more distilled and coherent final result, or we can think more in the terminology related to registered reports and replication, such as "the garden of forking paths" (Gelman & Loken, 2014) or "researcher degrees of freedom" (Simmons et al., 2011), where researchers have many opportunities throughout the research process to tweak their findings to reach statistical significance (or in the case of LGBQ-heterosexual disparities in the presence of mischievous responding, opportunities to avoid *losing* statistical significance). Indeed, the movement toward preregistered studies is driven by a goal to prespecify the details of the methods, to reduce the forking paths and degrees of freedom—all with the objective of *transparency in research* (Gehlbach & Robinson, 2018). In the case of potentially mischievous responders, there are a tremendous number of researcher degrees of freedom, from deciding whether to do anything at all about the issue of data validity, to which method(s) to use for detection, to which items (and response options) to include in the screener, and possibly, to which observations should be removed.

At this point in the field's understanding of the effects of mischievous responders, we would recommend a combination of preregistration and presenting a range of results under different assumptions. For example, if researchers have decided they want to address the issue of potentially mischievous responders using boosted regressions, then they can preregister the specifics of the boosted regression parameters (e.g., tuning, bagging, cross-validation stopping rules) and the items included in the boosted regression (e.g., height, carrot eating). We would caution, however, *against* prespecifying when to stop removing observations based on a rigid percentage of data (e.g., only 1% of data) or overly strict screening criteria (e.g., only removing cases if they provided *all* extreme responses to 10 screener items). Instead, we follow the recommendation of Cimpian et al. (2018) and recommend that researchers present a *range* of estimates based on different thresholds for screening out observations (e.g., estimates arrived at retaining all data, retaining 99%, 95%, 90%, and so on). Once mischievous responders are removed from the data, the estimated LGBQ-heterosexual disparities should converge to a relatively stable estimate (Cimpian et al., 2018; Robinson-Cimpian, 2014), but determining in advance when that stability will be achieved is futile. If the estimates do not converge, this is also useful information for consumers of researchers, so that they can assess the robustness of the findings on which they are basing education, health, and policy decisions. Thus, we would urge researchers to be more transparent in their dealings with issues of data validity, and this transparency may take the form of a *combination* of (1) preregistration to reduce researcher degrees of freedom in advance and (2) presenting a range of estimates that result from the remaining researcher degrees of freedom that could not be reasonably eliminated in advance.

Appendix A

Additional Details on Methods

Additional Details for Study 1. Using the 2015 YRBS, Cimpian et al. (2018) estimated LGBQ-heterosexual disparities for the 20 outcomes of interest using all of the data in their final analytic sample (72,641 males, 76,319 females), as well as provided an overall average disparity estimate across the 20 outcomes, with all analyses run separately by student reported sex (male or female). These disparity estimates provided a baseline estimate, or an estimate akin to assuming that there were no mischievous responders in the data.

Then, to identify potentially mischievous responders, Cimpian et al. (2018) invoked the assumption of no *differential* impact of their screener, which includes items related to the frequency that a student reports eating carrots, fruit, salad, and potatoes, whether they report having asthma, their most recent dentist visit, and their height. They then used a boosted regression (Friedman, 2001) to predict reporting

LGBQ identification from the screener items. Under the assumption of no differential impact of the screener, these screener items should not predict true LGBQ identification, but the screener items may help predict intentionally misreported LGBQ identification. For example, Cimpian et al. (2018) assumed that gay- and heterosexual-identified students should not differ in how often they ate carrots; yet, students who reported being "LGBQ" were much more likely to report eating carrots "4 or more times a day." Youth who reported being "LGBQ" were also far more likely to report eating salads, fruit, and potatoes with extreme frequency, to have missing values of height (importantly, the CDC recoded heights deemed "biologically implausible" to missing), to have never been to the dentist, and to be unsure if they had asthma.

The model $f(\cdot)$ predicting LGBQ identification to be discovered via boosted regression for individual respondent *i* is:

$$LGBQ_i = f\begin{pmatrix} screenerItem1_i, screenerItem2_i, \ldots, \\ screenerItem7_i, location_i, svywgt_i \end{pmatrix}$$

The specific R code—and thus, the tuning parameters—used to estimate this function is `gbm(sexmin ~ weight + q71 + q72 + q73 + q74 + q86 + q87 + stheight + location, data=renew, distribution = "bernoulli", n.trees = 10000, cv.folds=10, class.stratify.cv=TRUE, bag.fraction=0.5, train.fraction = 0.8, shrinkage = 0.01, interaction.depth = 3)`

LGBQ status (sexmin) is predicted by variables indicating consumption of fruit (q71 on the YRBS survey), salad (q72), potatoes (q73), and carrots (q74), frequency of dentist visits (q86), asthma (q87), height (variable: stheight), and location fixed effects. Sampling weights (variable: weight) are included in the models as predictors (see DuGoff et al., 2014, for more on the benefits of including weights in the estimation).

While this approach does require us to select the screener items included in the model, the functional form of the boosted regression is identified through an iterative process that does not rely on researcher assumptions regarding which outcome responses might be tempting for youth providing mischievous responses nor any assumptions regarding the relationships between items (Athey & Imbens, 2017; Friedman, 2001; McCaffrey et al., 2004; Mullainathan & Spiess, 2017). The regression generates a propensity (McCaffrey et al., 2004) for each respondent to report being LGBQ based on their responses to the screener items, which is then used as a proxy for likely mischievousness.

Estimating disparities. We use weighted linear probability models to obtain estimates, with ordinal values recoded as continuous, using location fixed effects:

$$Y_{ij} = \alpha_j + \beta LGBQ_{ij} + \varepsilon_{ij}$$

where outcome Y (e.g., sadness/hopelessness, suicide attempts) for respondent *i* in location *j* is predicted as a function of whether respondent *i* reports being LGBQ and *i*'s location, and α_j represents *J* location fixed effects. Because one of our goals is to compare disparity changes across models, we use linear probability models rather than logistic models, which are less suited to cross-model comparisons (Angrist & Pischke, 2009; Mood, 2010). The primary coefficient of interest is β, which represents the LGBQ-heterosexual disparity after accounting for location-level variation. All estimates are presented in standard deviation units, derived by dividing the unstandardized version of β by the standard deviation of Y. These estimates can then be interpreted as standardized differences (i.e., "effect sizes").

Confidence intervals. We estimate 95% asymmetric confidence intervals for each of the 20 outcomes examined via 1999 bootstrapped replications, including bootstrapping the boosted regressions, to account for measurement error in both the outcome estimation and propensity-score generation stages. As with the original YRBS sampling procedures, resampling occurs at both the location and primary sampling unit levels.

Average disparities. To obtain a stable, unbiased average estimate across the 20 outcomes, we use a precision-weighted random-effects model.

Differences between models in average disparities. We obtain the difference in estimates of average disparities between any two models *x* and *y* using simple subtraction, where each model includes the appropriate data set (i.e., model *x* uses all data, and model *y* includes all data minus the top XX% of suspected mischievous responders). Then, to test the statistical significance of these differences, we estimate the standard error using the following equation:

$$\sigma(\delta) = \left[\mathbf{X}'\Sigma^{-1}\mathbf{X}\right]^{-\frac{1}{2}}$$

where $\mathbf{X}$ is a 20 × 1 unit vector and $\boldsymbol{\Sigma}$ is a 20 × 20 matrix obtained from difference between model *x* and model *y* in each of the 1999 bootstrapped replications.

This method will lead to larger standard errors than those that would be obtained if the covariance of the outcomes was ignored (Schmelling, 1995); therefore, these tests are more conservative. This is particularly important as Cimpian and colleagues found statistically significant differences at each level of data removal for males, and frequently for females as well, indicating LGBQ-heterosexual disparities are sensitive to even a small subset of respondents providing mischievous responses.

Additional Details for Study 2. We use a random effects model to predict the *Reduction* in the estimate of LGBQ-heterosexual disparities between the model using all data and a given model with 1% to 25% of potential mischievous responders removed:

$$Reduction_{ij} = \alpha_{0j} + \alpha_{1j} lnPercentExtreme_{j} + \Gamma \mathbf{Model_{ij}} + \varepsilon_{ij} + \upsilon_{0j}$$

where *lnPercentExtreme* is the natural log (to account for a positive skew) of the percentage of males/females providing the most extreme response option in the full data set, and **Model** is a vector of 25 indicators of the estimate model (i.e., the model removing the top 1% of data, the top 2%, etc.). We also include random effects (υ_{0j}) to account for variability in average reductions across the 20 outcomes, and we cluster the robust standard errors on the 20 outcomes. Here, α_{1j} is the coefficient of interest, and we report standardized effect sizes.

Additional Details for Study 3

Method 2: Unconditional probability ranking. In this method, individual i 's value of P is the product of i 's response probabilities p for each item m in a group of items M (in our case, $M = 7$):

$$P_i = \prod_{m=1}^{M} p_{im}$$

Although p_{im} can be any kind of probability (e.g., a conditional probability), we follow Robinson-Cimpian's example and use a simple unconditional probability (i.e., the proportion of individuals who provided the response that individual i provided for item m). For example, in Table 4, if 1.05% of the respondents reported eating carrots "4 or more times a day," and 98.95% did not, then an individual who provided the low-frequency response of "4 or more times a day" would have $p_{im} = .0105$, and an individual who provided any other response would have $p_{im} = .9895$ for that item. The more prespecified low-frequency items selected by an individual, the lower that individual's product will be.

Once the index of P is created and respondents are ranked, we estimate a series of LGBQ-heterosexual disparities just as we did for the boosted regression method. That is, we estimate the disparities first using the full analytic data set, then again after removing the top 1% of cases with the highest mischievousness index, then again removing another 1%, and so on.

Method 3: Count based. The count-based removal of suspected mischievous responders is similar to the probability-based approach, except each prespecified low-frequency response-option receives equal weight. Referring to Table 4, the response-option of "not sure" if you have asthma would receive less weight than eating carrots "4 or more times a day" in the probability-based approach (because the carrot response is lower frequency), but would receive equal weight in the count-based approach. More simply, this approach just tallies up the number of low-frequency responses provided to the items in Table 4. The more low-frequency response-options a respondent provides, the more likely they are to be mischievous.

This method is the most commonly used screening technique in the literature (e.g., Furlong et al., 2004; Furlong et al., 2017; Mittleman, 2018; Robinson-Cimpian, 2014). In some instances, it is applied implicitly, such as in Mittleman (2018) who removed observations only if they respond affirmatively to all screener items, or in Robinson and Espelage (2011) who removed observations if they responded affirmatively to two or more screener items; this also illustrates the degree of flexibility researchers have in deciding how strict or lenient they will be when screening out suspicious observations, which is common to Methods 1 to 3.

As with other methods, we estimate the disparities with the full data set, then remove the observations providing the top 1% of extreme responses (which includes individuals providing *all* low-frequency response options on the screener) and reestimate the disparities, then remove observations providing all but one low-frequency response, and so on.

Method 4: Regression adjustment. The regression adjustment approach uses the values of *P* from the probability-based approach, but statistically conditions on a function of *P* rather than remove observations based on the ranking of *P*. Robinson-Cimpian (2014) proposed this approach as a possible method for addressing mischievous responders and applied it as a supplemental analysis to data from the 2012 Dane County Youth Assessment. Fish and Russell (2018) recently applied this method to the Add Health data set.

The consequences of different functional forms of *P* (e.g., logging *P*) on how the estimates of the disparities are altered will also be examined. Specifically, we estimate the following 3 functional forms:

$$\text{Linear: } Y_i = \beta_0 + \beta_1 LGBQ_i + \beta_2 P_i + \varepsilon_i$$

$$\text{Natural logged:} Y_i = \beta_0 + \beta_1 LGBQ_i + \beta_2 ln\left(P_i\right) + \varepsilon_i$$

$$\text{Quartic of natural log: } Y_i = \beta_0 + \beta_1 LGBQ_i + \beta_2 ln\left(P_i\right) + \beta_3 \left(ln\left(P_i\right)\right)^2 + \beta_4 \left(ln\left(P_i\right)\right)^3 + \beta_5 \left(ln\left(P_i\right)\right)^4 + \varepsilon_i$$

where *Y* is an outcome for individual *i, P* is the probability-based index of mischievousness, and β_1 is the coefficient of interest, as it represents the LGBQ-heterosexual disparity. These functional forms were selected because they may be intuitively used (i.e., the linear) or they have been applied in the existing literature (i.e., the natural log by Fish & Russell, 2018; the quartic of the natural log by Robinson-Cimpian, 2014).

Appendix B

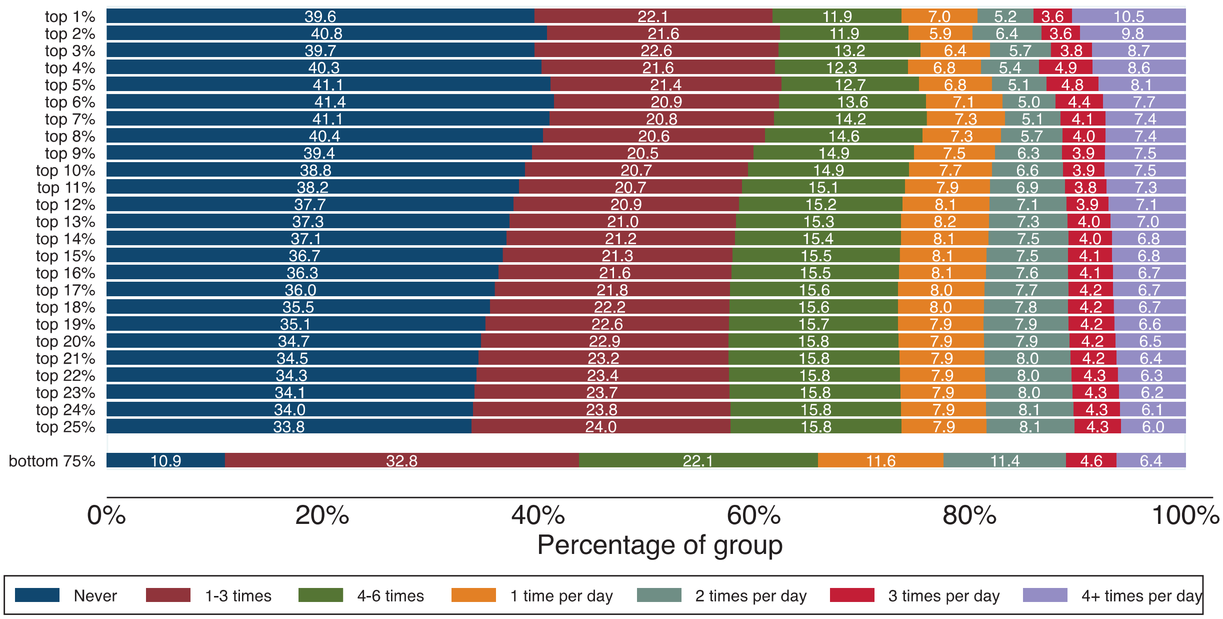


FIGURE B1. *Distribution of responses to fruit item for males, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

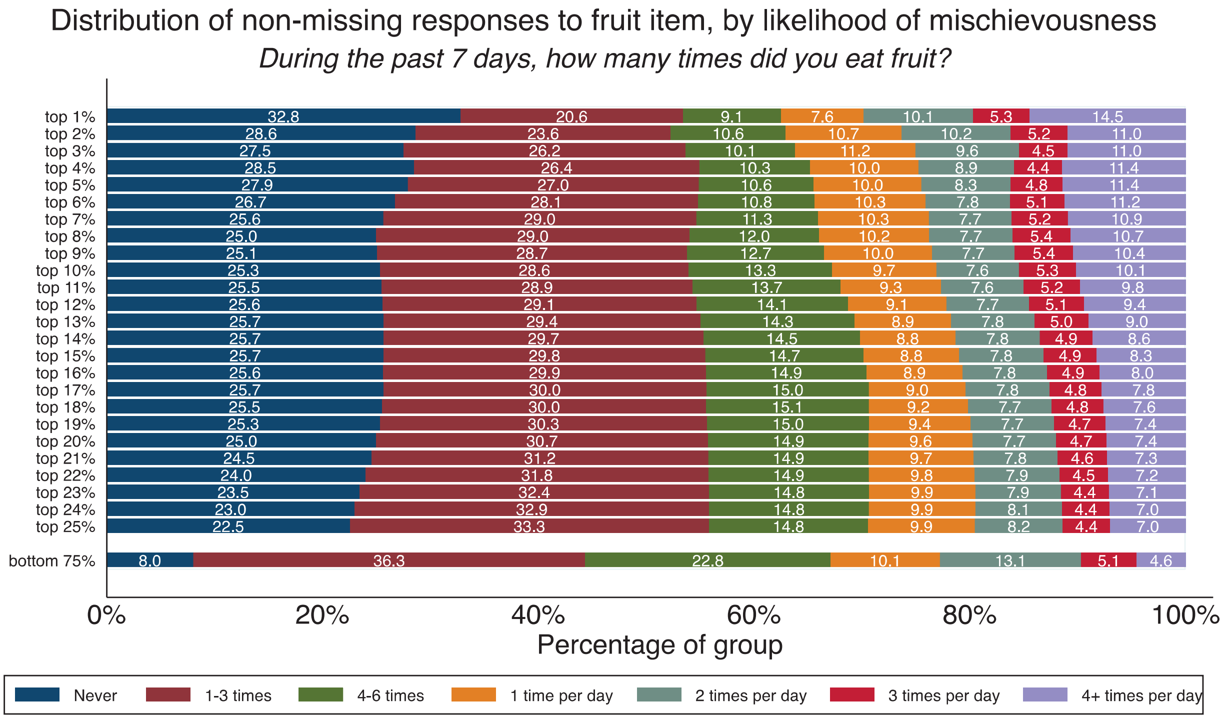


FIGURE B2. *Distribution of responses to fruit item for females, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

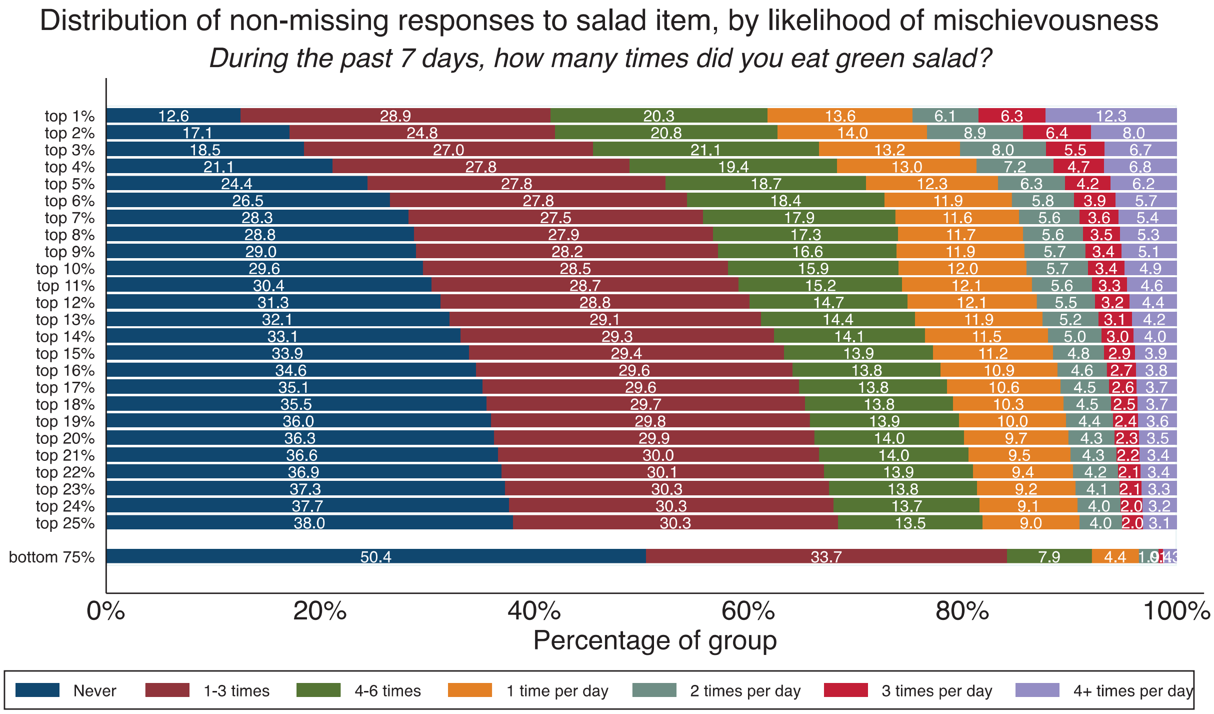


FIGURE B3. *Distribution of responses to salad item for males, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

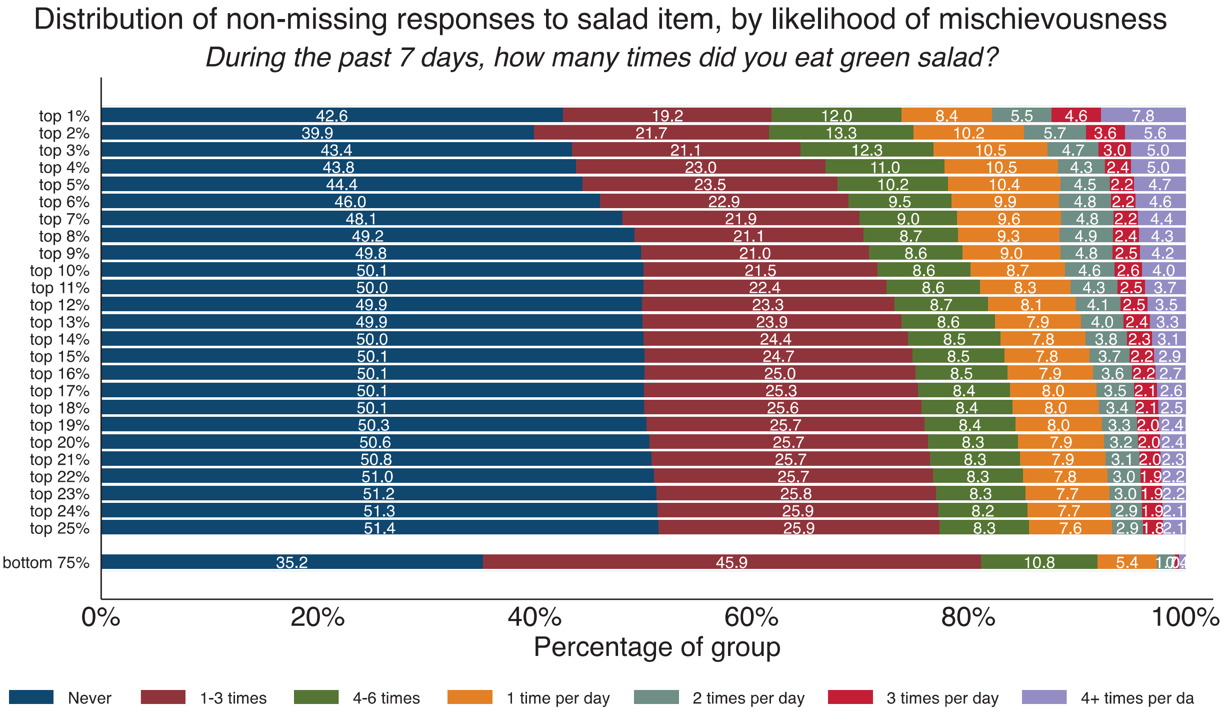


FIGURE B4. *Distribution of responses to salad item for females, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

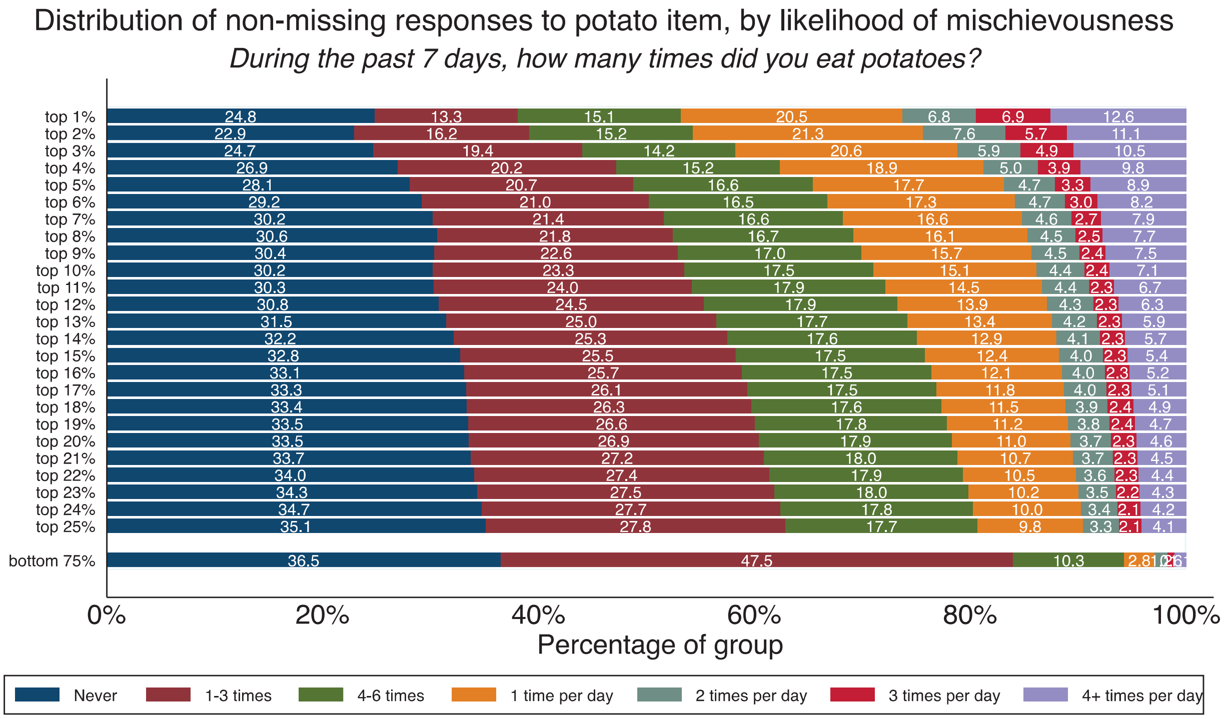


FIGURE B5. *Distribution of potato consumption for males, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

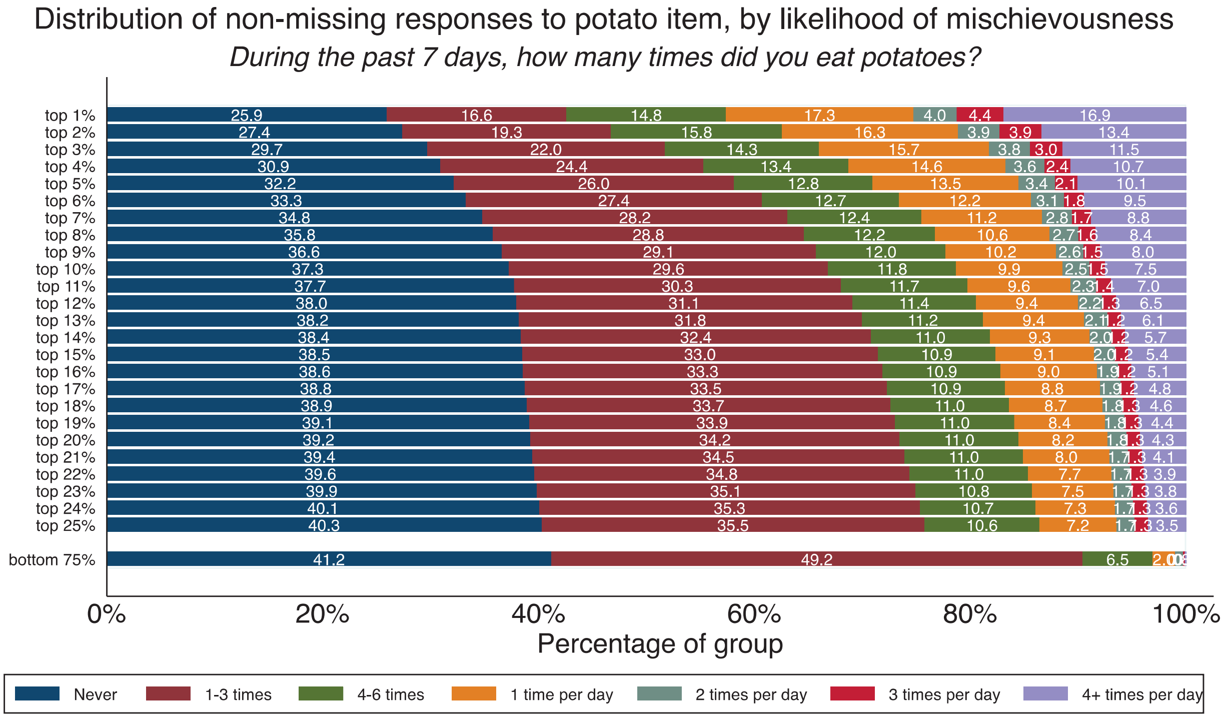


FIGURE B6. *Distribution of potato consumption for females, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

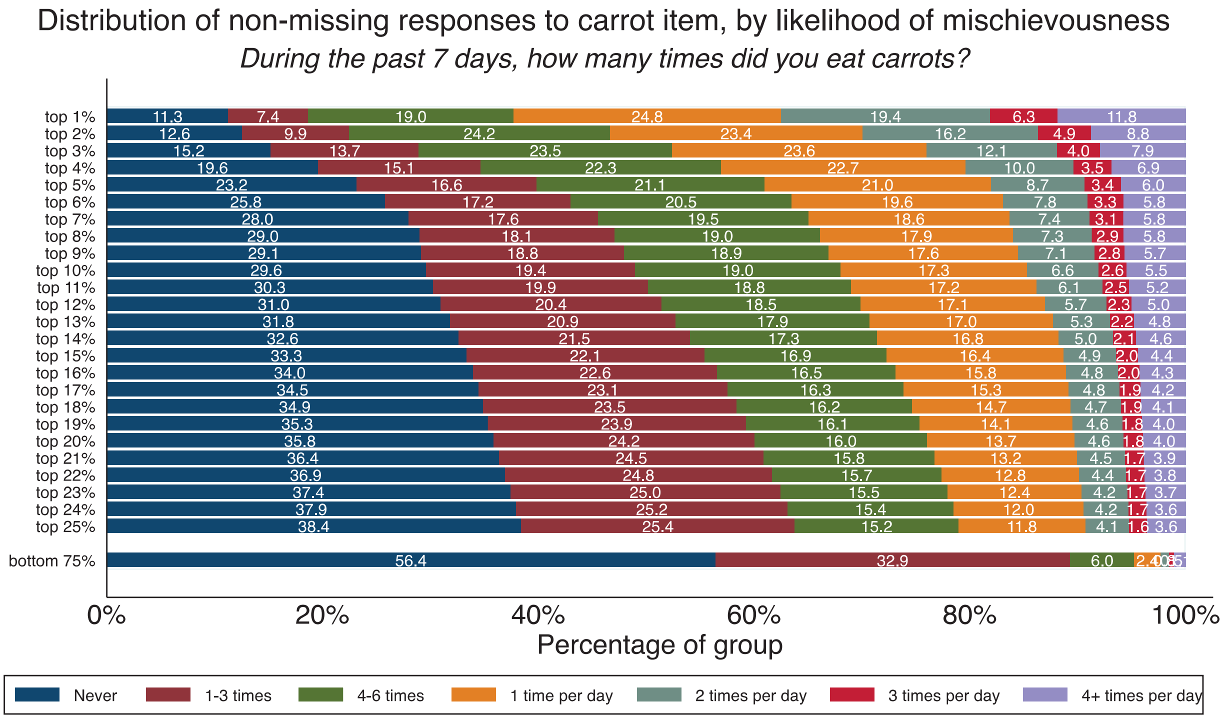


FIGURE B7. *Distribution of carrot consumption for males, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

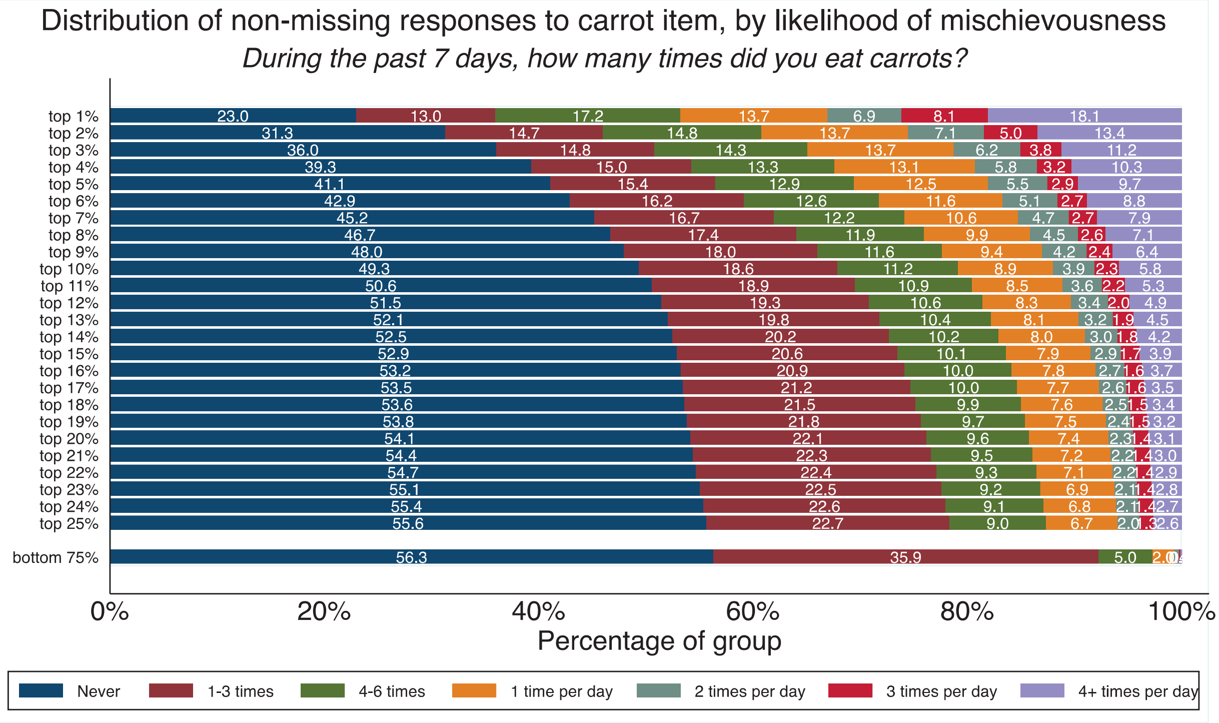


FIGURE B8. *Distribution of carrot consumption for females, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

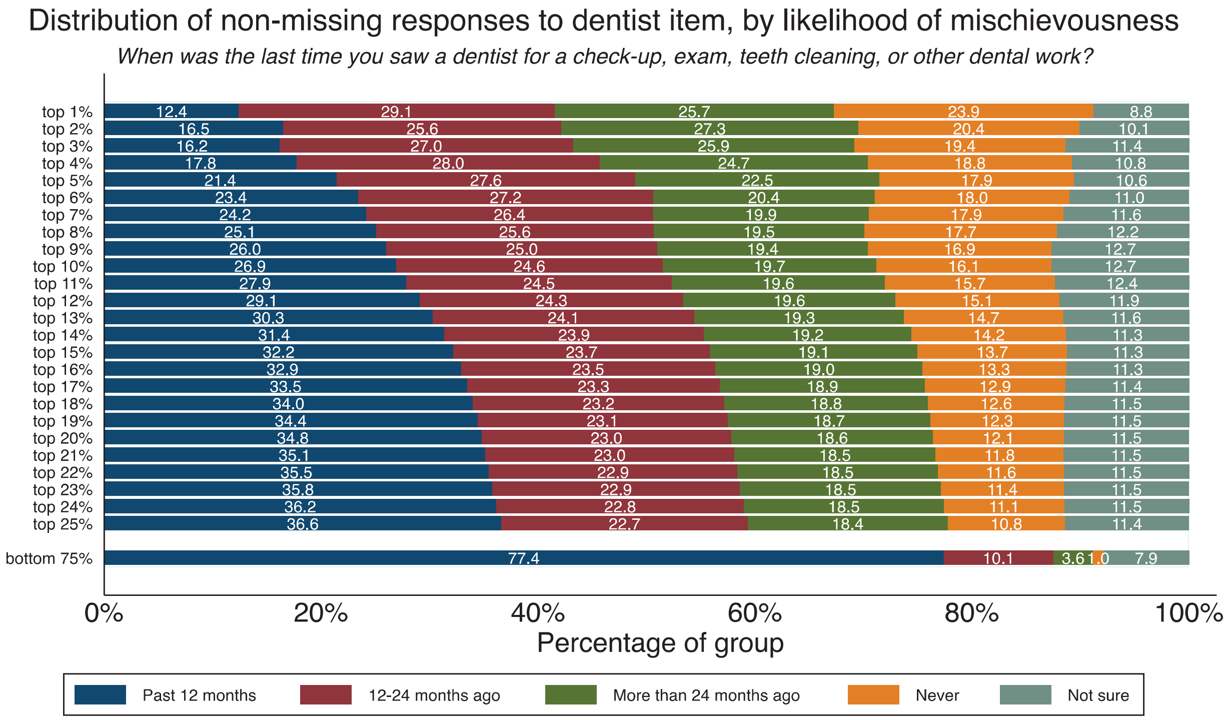


FIGURE B9. *Distribution of dentist visits for males, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

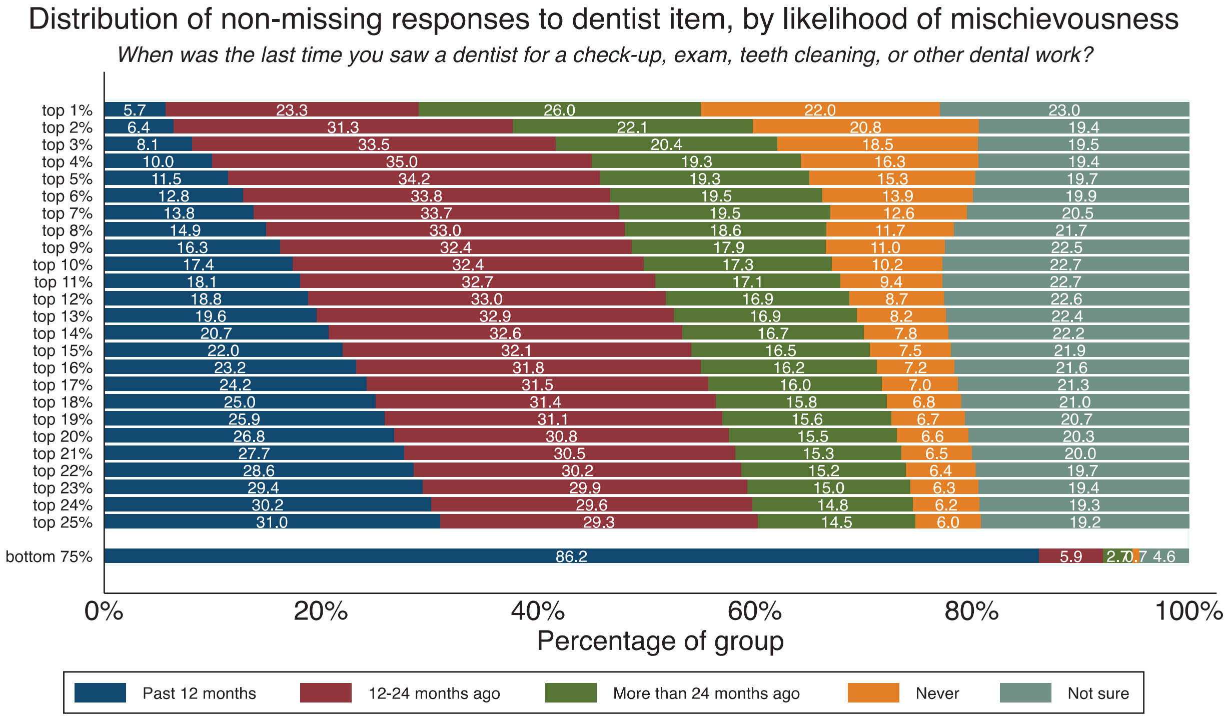


FIGURE B10. *Distribution of dentist visits for females, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

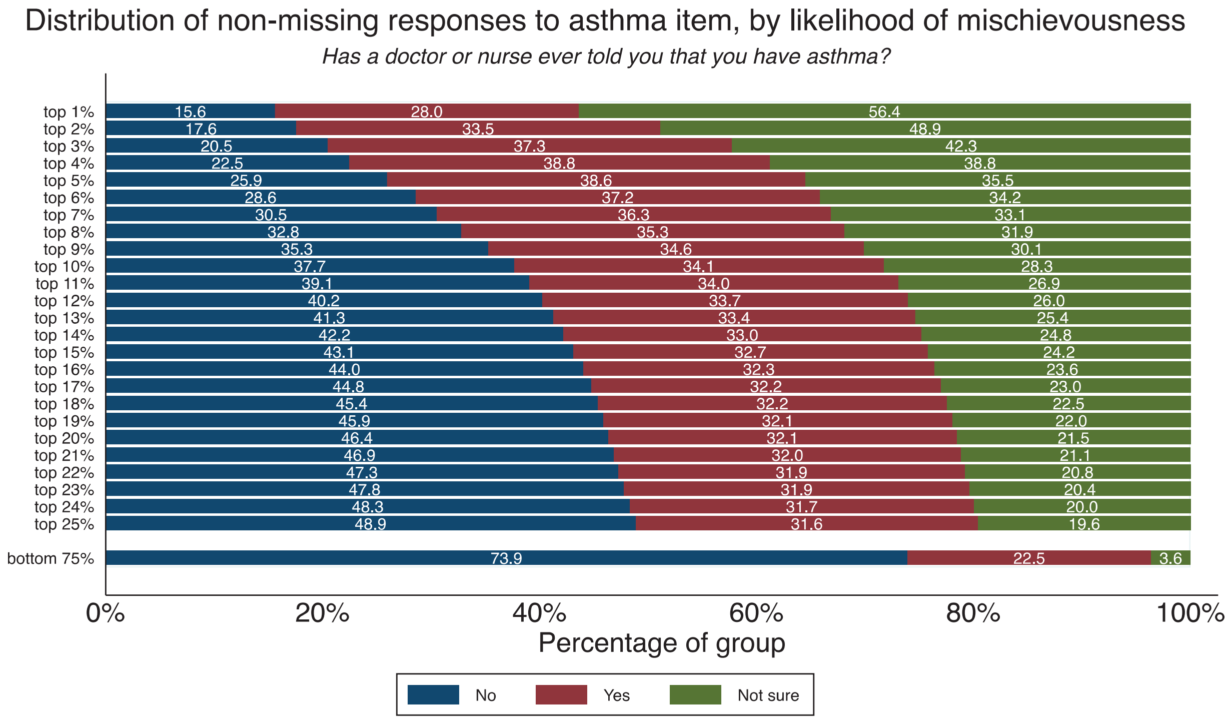


FIGURE B11. *Distribution of asthma diagnosis responses for males, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

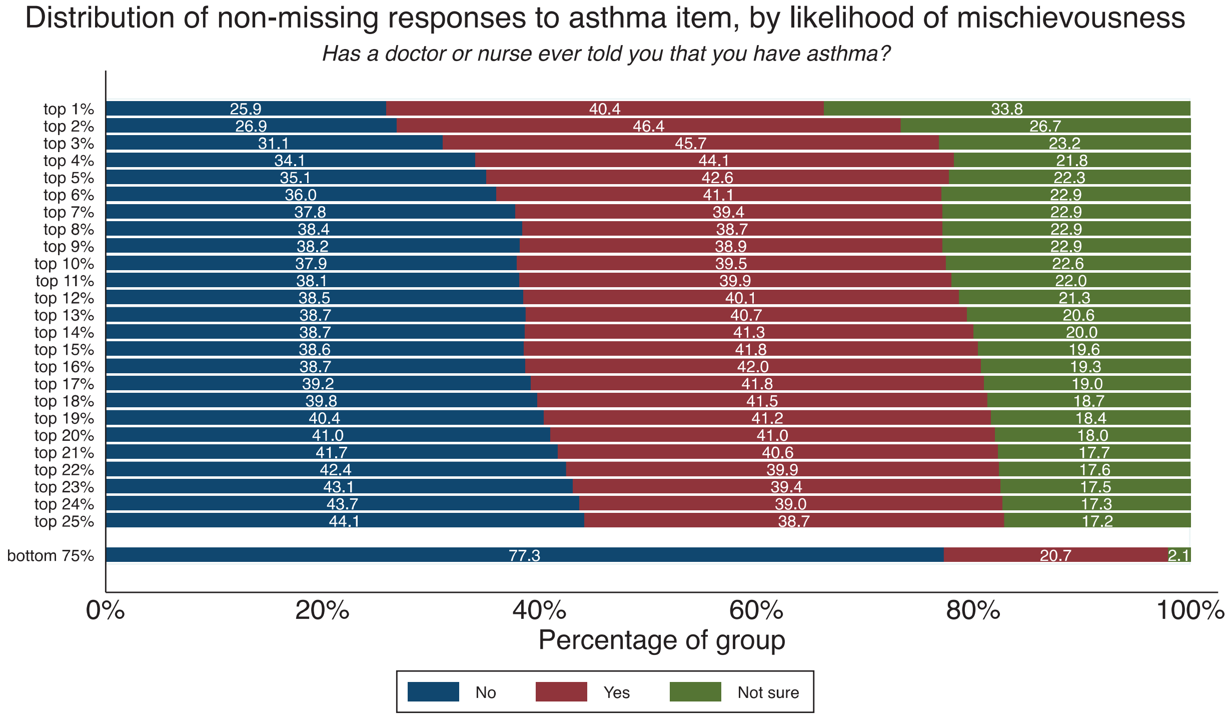


FIGURE B12. *Distribution of asthma diagnosis responses for females, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

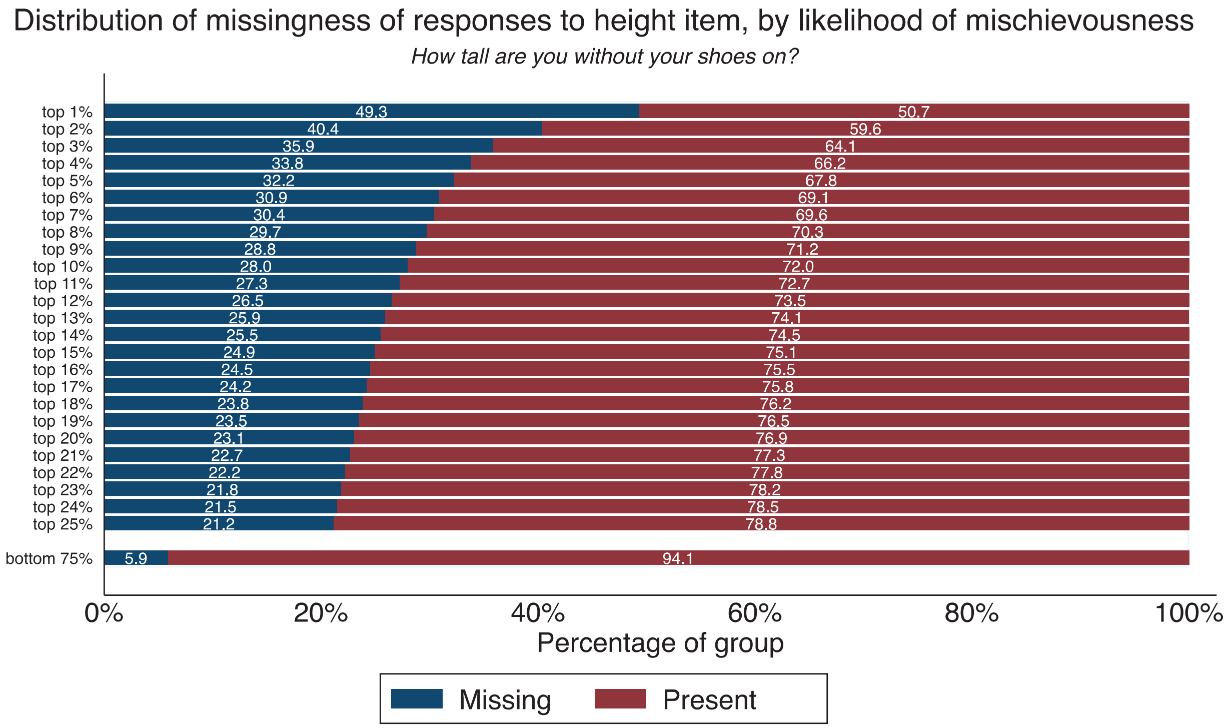


FIGURE B13. *Distribution of height missingness for males, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

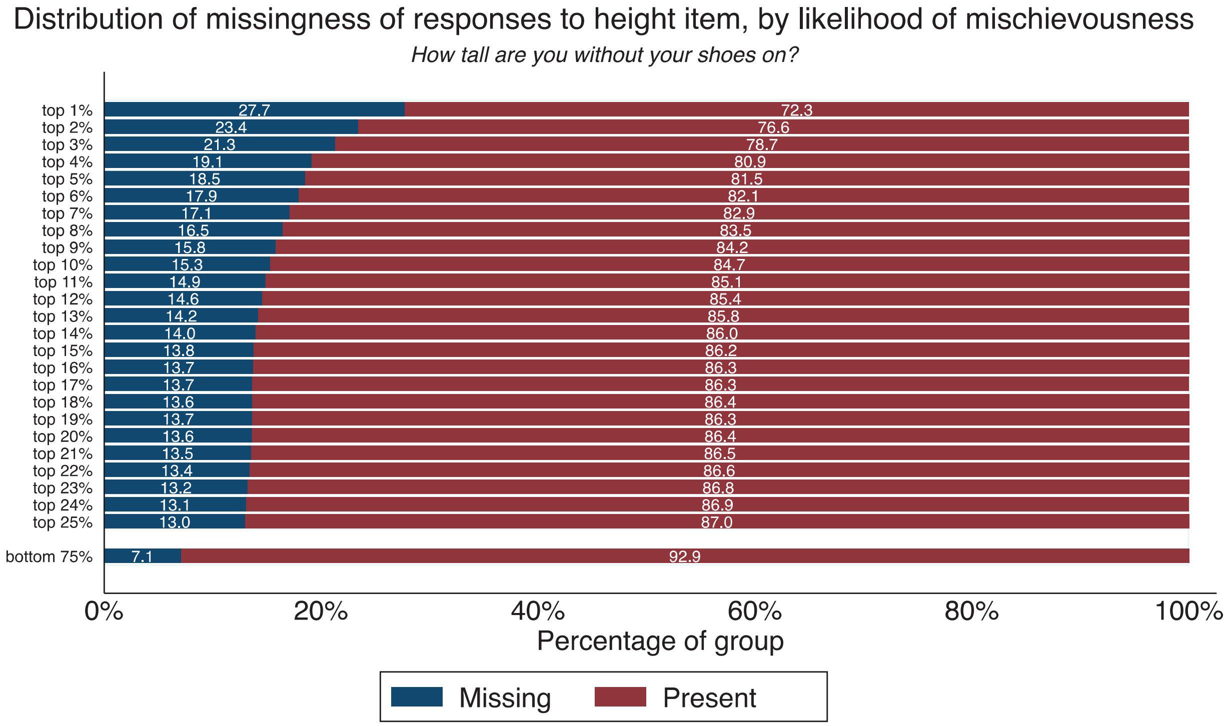


FIGURE B14. *Distribution of height missingness for females, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

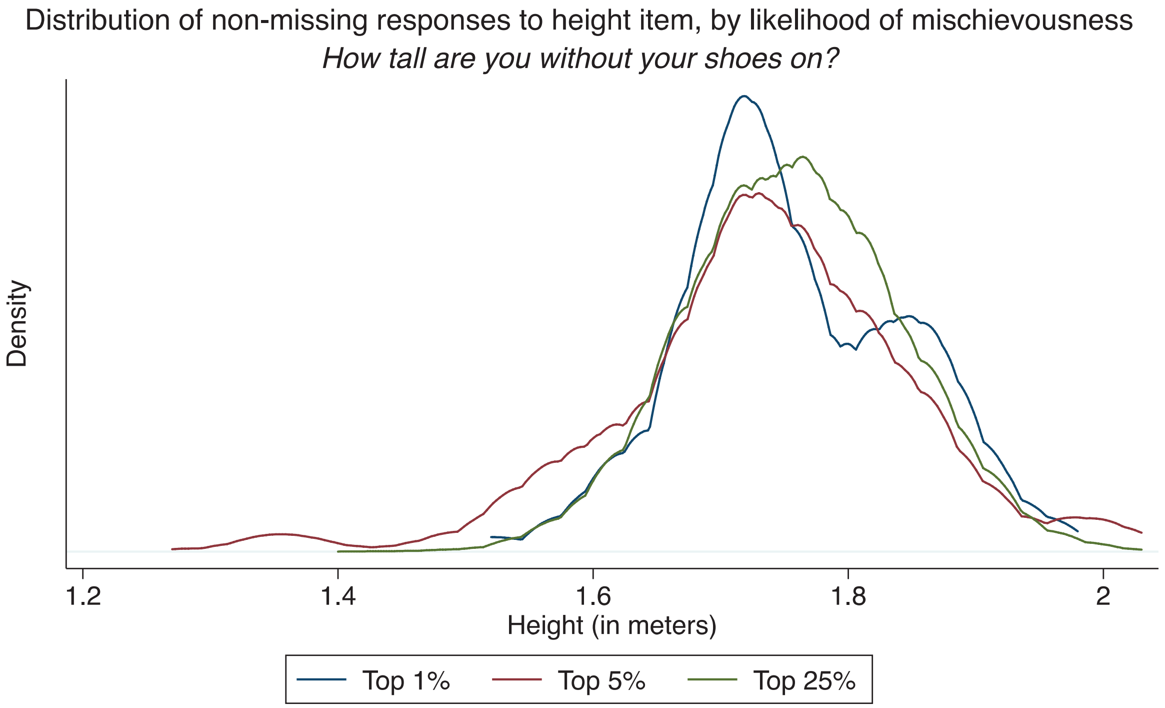


FIGURE B15. *Distribution of nonmissing height for males, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*

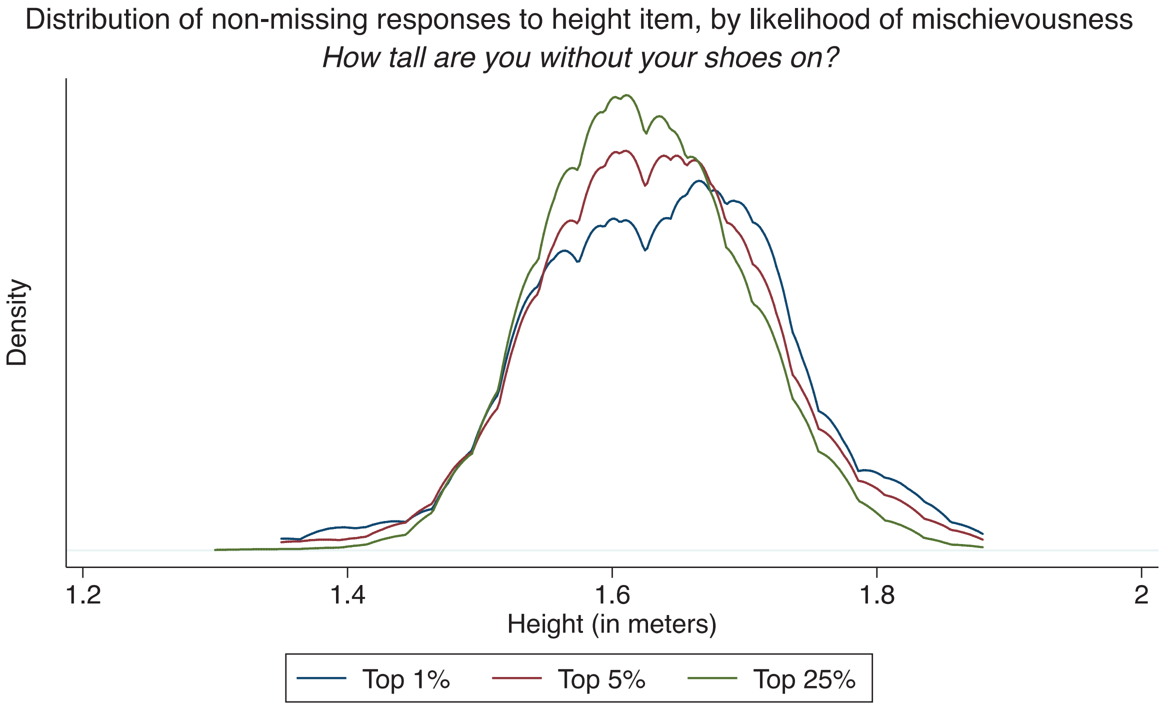
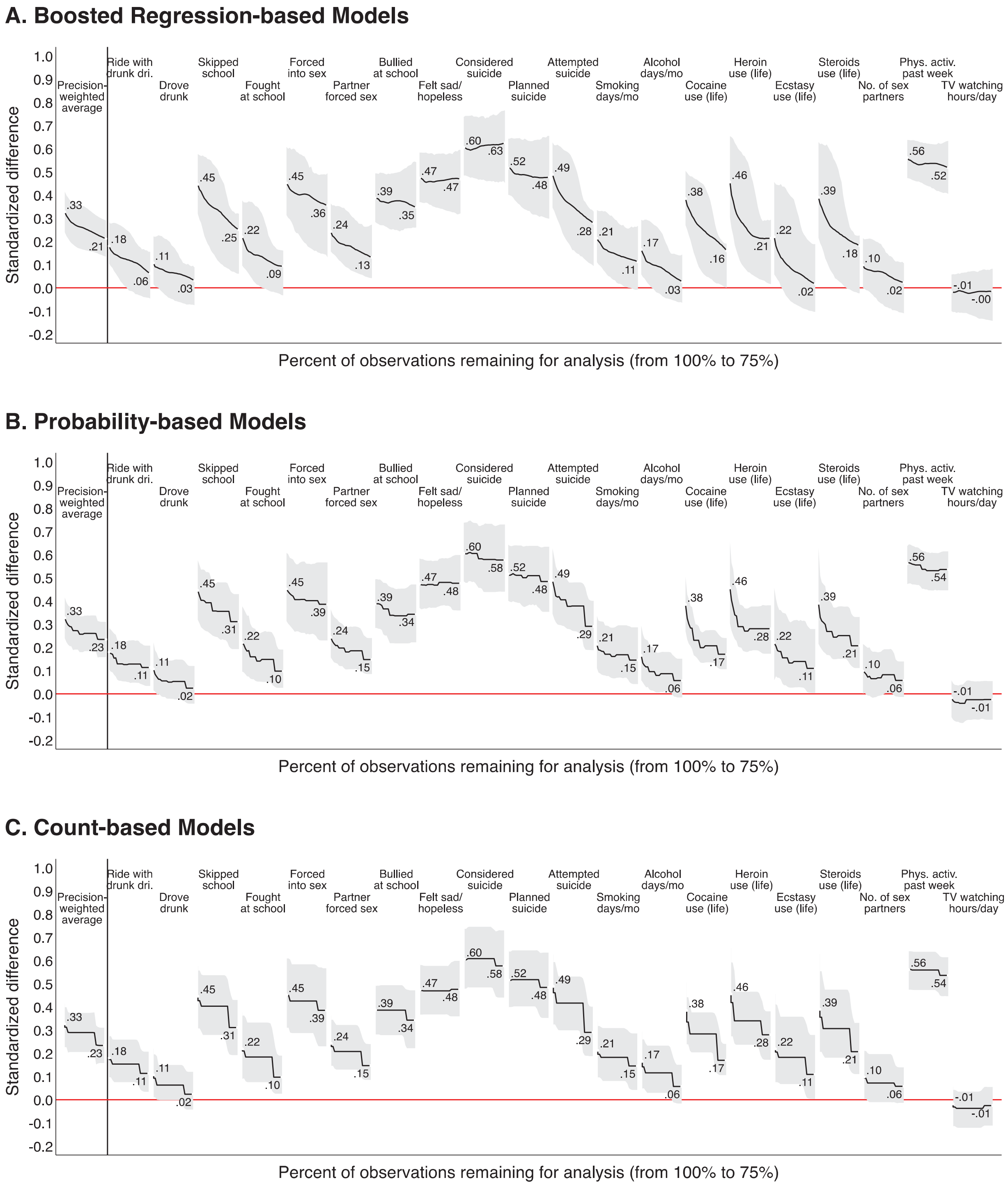


FIGURE B16. *Distribution of nonmissing height for females, subset of Youth Risk Behavior Survey (YRBS) 2017 who were administered all seven screener items.*



A. Boosted Regression-based Models

B. Probability-based Models

C. Count-based Models

FIGURE B17. *Average LGBQ-Heterosexual disparity among reported males in the full sample, by model, outcome, and percent of observations screened out.*

Note. LGBQ = lesbian, gay, bisexual, or questioning. The shaded areas represent asymmetrical 95% confidence intervals (*CIs*), constructed via 1999 bootstrapped samples. If the shaded area does not cross the horizontal red line at zero, the disparity is statistically significant ($p < .05$).

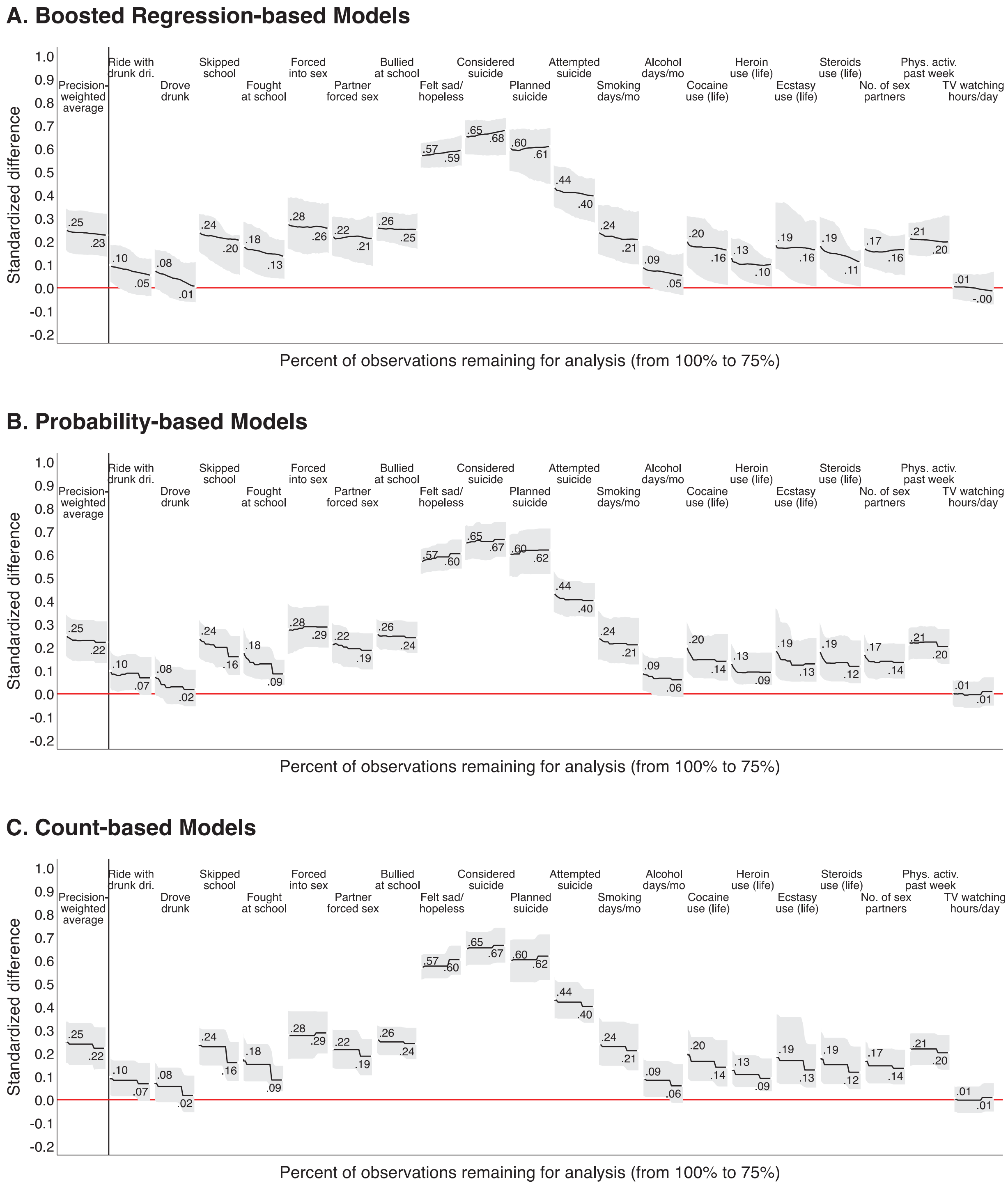


FIGURE B18. *Average LGBQ-Heterosexual disparity among reported females in the full sample, by model, outcome, and percentage of observations screened out.*

Note. LGBQ = lesbian, gay, bisexual, or questioning. The shaded areas represent asymmetrical 95% CIs, constructed via *1999* bootstrapped samples. If the shaded area does not cross the horizontal red line at zero, the disparity is statistically significant ($p < .05$).

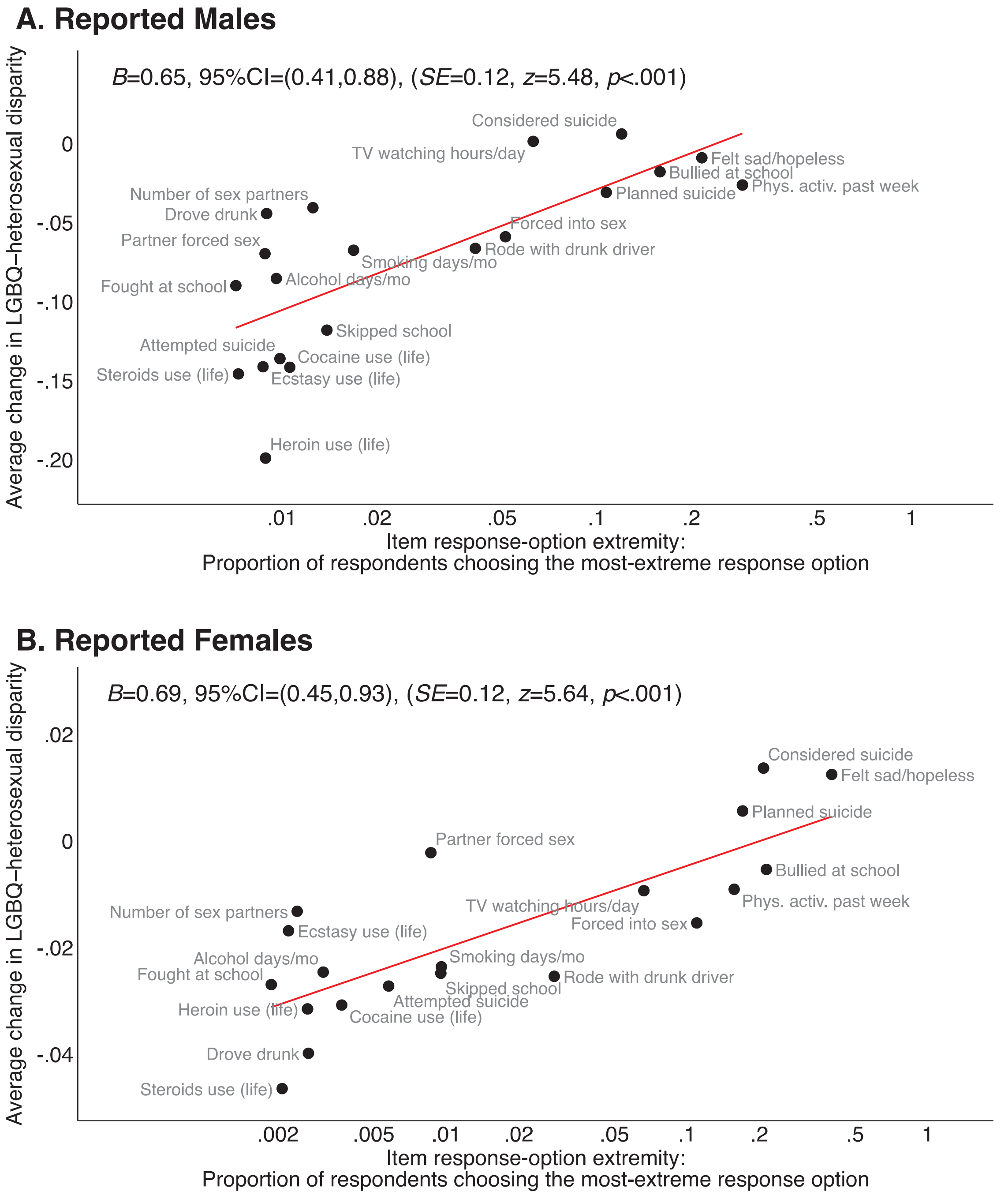


FIGURE B19. *Relationship between how much an item-level disparity changed when screening mischievous responders and the item response-option extremity, in the full sample.*

Note. More *extreme* item response-options (to the left on the *x*-axis) mean fewer respondents chose the most extreme option (e.g., using heroin "40 or more times"), which corresponds to larger average changes in the disparities when screening out mischievous responders.

Appendix C

2017 YRBS questionnaire instructions and relevant items.
Source: https://www.cdc.gov/healthyyouth/data/yrbs/questionnaires.htm

2017 State and Local Youth Risk Behavior Survey

This survey is about health behavior. It has been developed so you can tell us what you do that may affect your health. The information you give will be used to improve health education for young people like yourself.

DO NOT write your name on this survey. The answers you give will be kept private. No one will know what you write. Answer the questions based on what you really do.

Completing the survey is voluntary. Whether or not you answer the questions will not affect your grade in this class. If you are not comfortable answering a question, just leave it blank.

The questions that ask about your background will be used only to describe the types of students completing this survey. The information will not be used to find out your name. No names will ever be reported.

Make sure to read every question. Fill in the ovals completely. When you are finished, follow the instructions of the person giving you the survey.

Thank you very much for your help.

Directions
Use a #2 pencil only.
Make dark marks.
Fill in a response like this: A B D.
If you change your answer, erase your old answer completely.

Items Used for Screener

6. How tall are you without your shoes on?
Directions: Write your height in the shaded blank boxes. Fill in the matching oval below each number.

71. During the past 7 days, how many times did you eat **fruit**? (Do **not** count fruit juice.)
A. I did not eat fruit during the past 7 days
B. 1 to 3 times during the past 7 days
C. 4 to 6 times during the past 7 days
D. 1 time per day
E. 2 times per day
F. 3 times per day
G. 4 or more times per day

72. During the past 7 days, how many times did you eat **green salad**?
A. I did not eat green salad during the past 7 days
B. 1 to 3 times during the past 7 days
C. 4 to 6 times during the past 7 days
D. 1 time per day
E. 2 times per day
F. 3 times per day
G. 4 or more times per day

73. During the past 7 days, how many times did you eat **potatoes**? (Do **not** count french fries, fried potatoes, or potato chips.)
A. I did not eat potatoes during the past 7 days
B. 1 to 3 times during the past 7 days
C. 4 to 6 times during the past 7 days
D. 1 time per day
E. 2 times per day
F. 3 times per day
G. 4 or more times per day

74. During the past 7 days, how many times did you eat **carrots**?
A. I did not eat carrots during the past 7 days
B. 1 to 3 times during the past 7 days
C. 4 to 6 times during the past 7 days
D. 1 time per day
E. 2 times per day
F. 3 times per day
G. 4 or more times per day

86. When was the last time you saw a dentist for a check-up, exam, teeth cleaning, or other dental work?
A. During the past 12 months
B. Between 12 and 24 months ago
C. More than 24 months ago
D. Never
E. Not sure

87. Has a doctor or nurse ever told you that you have asthma?
A. Yes
B. No
C. Not sure

Items Used as Outcomes

9. During the past 30 days, how many times did you **ride** in a car or other vehicle **driven by someone who had been drinking alcohol**?
   - A. 0 times
   - B. 1 time
   - C. 2 or 3 times
   - D. 4 or 5 times
   - E. 6 or more times

10. During the past 30 days, how many times did you **drive** a car or other vehicle **when you had been drinking alcohol**?
   - A. I did not drive a car or other vehicle during the past 30 days
   - B. 0 times
   - C. 1 time
   - D. 2 or 3 times
   - E. 4 or 5 times
   - F. 6 or more times

15. During the past 30 days, on how many days did you **not** go to school because you felt you would be unsafe at school or on your way to or from school?
   - A. 0 days
   - B. 1 day
   - C. 2 or 3 days
   - D. 4 or 5 days
   - E. 6 or more days

18. During the past 12 months, how many times were you in a **physical fight on school property**?
   - A. 0 times
   - B. 1 time
   - C. 2 or 3 times
   - D. 4 or 5 times
   - E. 6 or 7 times
   - F. 8 or 9 times
   - G. 10 or 11 times
   - H. 12 or more times

19. Have you ever been physically forced to have sexual intercourse when you did not want to?
   - A. Yes
   - B. No

21. During the past 12 months, how many times did **someone you were dating or going out with** force you to do sexual things that you did not want to do? (Count such things as kissing, touching, or being physically forced to have sexual intercourse.)
   - A. I did not date or go out with anyone during the past 12 months
   - B. 0 times
   - C. 1 time
   - D. 2 or 3 times
   - E. 4 or 5 times
   - F. 6 or more times

23. During the past 12 months, have you ever been bullied **on school property**?
   - A. Yes
   - B. No

25. During the past 12 months, did you ever feel so sad or hopeless almost every day for **two weeks or more in a row** that you stopped doing some usual activities?
   - A. Yes
   - B. No

26. During the past 12 months, did you ever **seriously** consider attempting suicide?
   - A. Yes
   - B. No

27. During the past 12 months, did you make a plan about how you would attempt suicide?
   - A. Yes
   - B. No

28. During the past 12 months, how many times did you actually attempt suicide?
   - A. 0 times
   - B. 1 time
   - C. 2 or 3 times
   - D. 4 or 5 times
   - E. 6 or more times

32. During the past 30 days, on how many days did you smoke cigarettes?
   - A. 0 days
   - B. 1 or 2 days
   - C. 3 to 5 days
   - D. 6 to 9 days
   - E. 10 to 19 days
   - F. 20 to 29 days
   - G. All 30 days

42. During the past 30 days, on how many days did you have at least one drink of alcohol?
   - A. 0 days
   - B. 1 or 2 days
   - C. 3 to 5 days
   - D. 6 to 9 days
   - E. 10 to 19 days
   - F. 20 to 29 days
   - G. All 30 days

49. During your life, how many times have you used **any** form of cocaine, including powder, crack, or freebase?
    A. 0 times
    B. 1 or 2 times
    C. 3 to 9 times
    D. 10 to 19 times
    E. 20 to 39 times
    F. 40 or more times

51. During your life, how many times have you used **heroin** (also called smack, junk, or China White)?
    A. 0 times
    B. 1 or 2 times
    C. 3 to 9 times
    D. 10 to 19 times
    E. 20 to 39 times
    F. 40 or more times

53. During your life, how many times have you used **ecstasy** (also called MDMA)?
    A. 0 times
    B. 1 or 2 times
    C. 3 to 9 times
    D. 10 to 19 times
    E. 20 to 39 times
    F. 40 or more times

55. During your life, how many times have you taken **steroid pills or shots** without a doctor's prescription?
    A. 0 times
    B. 1 or 2 times
    C. 3 to 9 times
    D. 10 to 19 times
    E. 20 to 39 times
    F. 40 or more times

61. During your life, with how many people have you had sexual intercourse?
    A. I have never had sexual intercourse
    B. 1 person
    C. 2 people
    D. 3 people
    E. 4 people
    F. 5 people
    G. 6 or more people

79. During the past 7 days, on how many days were you physically active for a total of **at least 60 minutes per day**? (Add up all the time you spent in any kind of physical activity that increased your heart rate and made you breathe hard some of the time.)
    A. 0 days
    B. 1 day
    C. 2 days
    D. 3 days
    E. 4 days
    F. 5 days
    G. 6 days
    H. 7 days

80. On an average school day, how many hours do you watch TV?
    A. I do not watch TV on an average school day
    B. Less than 1 hour per day
    C. 1 hour per day
    D. 2 hours per day
    E. 3 hours per day
    F. 4 hours per day
    G. 5 or more hours per day

Other Relevant Survey Items

[Note: We separate analyses by sex (item 2). The sexual identity question (item 67) is used to determine LGBQ identification. We use the other items (1, 3, and 5) to see how screening affects the other demographic variables, thereby providing an indication of how screening affects generalizability.]

1. How old are you?
    A. 12 years old or younger
    B. 13 years old
    C. 14 years old
    D. 15 years old
    E. 16 years old
    F. 17 years old
    G. 18 years old or older

2. What is your sex?
    A. Female
    B. Male

3. In what grade are you?
    A. 9th grade
    B. 10th grade
    C. 11th grade
    D. 12th grade
    E. Ungraded or other grade

5. What is your race? **(Select one or more responses.)**
    A. American Indian or Alaska Native
    B. Asian
    C. Black or African American
    D. Native Hawaiian or Other Pacific Islander
    E. White

67. Which of the following best describes you?
    A. Heterosexual (straight)
    B. Gay or lesbian
    C. Bisexual
    D. Not sure

ORCID iD

Joseph R. Cimpian https://orcid.org/0000-0001-6111-8895

Data Availability

To facilitate the use of screening analyses with the 2017 YRBS data, as well as to further increase the transparency of our own research, we make our main screening weights and code freely available for download from the Open ICPSR website (https://www.openicpsr.org/openicpsr/project/115086/version/V1/view/). Thus, researchers can readily assess the robustness of their own findings with the 2017 YRBS data, as well create their own screener with the 2017 YRBS or use our code as a template for a different data set.

Notes

1. There are exceptions, of course, such as Mittleman's (2018) recent use of the Fragile Families data set to examine differential disciplinary practices by sexual identity. The structure of that data set is more similar to the Add Health data set—which has contributed substantially to the LGBQ student literature, though its validity has recently come into question (Savin-Williams & Joyner, 2014a, 2014b; but cf. Fish & Russell, 2018; Katz-Wise et al., 2015; Li et al., 2014)—with nonanonymized longitudinal data, different modes of survey administration, and questions of relatives and educators. It is worth noting, though, that despite the different nature of Mittleman's data, he too uses a screener, suggesting the growing importance of validity screening with LGBQ status data across data sets.

2. Note that these extreme response options are based on the pattern of results from Cimpian et al. (2018), which used a boosted regression to identify likely mischievous responders. Thus, all of the other methods (i.e., Methods 2, 3, and 4) benefit from this knowledge, and consequently, perform better than they would if they did not have this knowledge from the boosted regression results in the study we are replicating. That is, if anything, the comparison of our methods is likely biased toward zero.

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Badgett, L. (2009). *Best practices for asking questions about sexual orientation on surveys*. Los Angeles, CA: Williams Institute.

Brockenbrough, E. (2017). Outing the politics of knowledge production: A review of *LGBTQ Issues in Education: Advancing a Research Agenda. Educational Researcher*, *46*, 548–550.

Centers for Disease Control and Prevention. (2016). *MMWR* Surveillance Summaries. *MMWR Morbidity and Mortality Weekly Report, 65*. Retreived from https://www.cdc.gov/mmwr/indss_2016.html

Centers for Disease Control and Prevention. (2017). *YRBS journal articles by DASH authors*. Retrieved from https://www.cdc.gov/healthyyouth/data/yrbs/pdf/yrbs_journal_articles_v3.pdf

Centers for Disease Control and Prevention. (2018). *2017 YRBS data user's guide*. Retrieved from https://www.cdc.gov/healthyyouth/data/yrbs/pdf/2017/2017_YRBS_Data_Users_Guide.pdf

Cimpian, J. R. (2017). Classification errors and bias regarding research on sexual minority youths. *Educational Researcher*, *46*, 517–529.

Cimpian, J. R., & Herrington, C. D. (2017). Editors' introduction: Introducing a methodological research collection on pressing issues for LGBTQ students. *Educational Researcher*, 46, 495–497.

Cimpian, J. R., Timmer, J. D., Birkett, M., Marro, R., Turner, B., & Phillips, G. L. (2018). Bias from potentially mischievous responders on large-scale estimates of LGBQ-heterosexual youth health disparities. *American Journal of Public Health, 108*, S258–S265. doi:10.2105/AJPH.2018.304407

Clayton, H. B., Lowry, R., August, E., & Jones, S. E. (2016). Nonmedical use of prescription drugs and sexual risk behaviors. *Pediatrics*, *137*, e20152480.

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

Cornell, D. G., & Loper, A. B. (1998). Assessment of violence and other high-risk behaviors with a school survey. *School Psychology Review*, *27*, 317–330.

Cross, J. E., & Newman-Gonchar, R. (2004). Data quality in student risk behavior surveys and administrator training. *Journal of School Violence*, *3*, 89–108.

DuGoff, E. H., Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results applying propensity score methods to complex surveys. *Health Services Research*, *49*, 284–303.

Dynarski, S., & Berends, M. (2015). Introduction to special issue. *Educational Evaluation and Policy Analysis, 37*(1 Suppl.), 3S–5S.

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

Fan, X., Miller, B. C., Park, K.-E., Winward, B. W., Christensen, M., Grotevant, H. D., & Tai, R. H. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods*, *18*, 223–244.

Fanelli, D., & Ioannidis, J. P. (2013). US studies may overestimate effect sizes in softer research. *Procedures of the National Academy of Sciences of the U S A*, *110*, 15031–15036.

Fish, J. N., & Russell, S. T. (2018). Have mischievous responders misidentified sexual minority youth disparities in the National Longitudinal Study of Adolescent to Adult Health? *Archives of Sexual Behavior*, *47*, 1053–1067.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232.

Furlong, M. J., Fullchange, A., & Dowdy, E. (2017). Effects of mischievous responding on universal mental health screening: I love rum raisin ice cream. *School Psychology Quarterly*, *32*, 320–335.

Furlong, M. J., Sharkey, J. D., Bates, M. P., & Smith, D. C. (2004). An examination of the reliability, data screening procedures, and extreme response patterns for the Youth Risk Behavior Surveillance Survey. *Journal of School Violence*, *3*, 109–130.

Gehlbach, H., & Robinson, C. D. (2018). Mitigating illusory results through preregistration in education. *Journal of Research on Educational Effectiveness*, *11*, 296–315.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465.

Groves, R. M., Jr., Fowler, F. J., Couper, M. P., Lepkowski, J. M., & Tourangeau, R. (2011). *Survey methodology*. Hoboken, NJ: John Wiley.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer. Retrieved from https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf (PDF with corrections as of January 2017)

Jia, Y., Konold, T. R., Cornell, D., & Huang, F. (2018). The impact of validity screening on associations between self-reports of bullying victimization and student outcomes. *Educational and Psychological Measurement*, *78*, 80–102.

Katz-Wise, S. L., Calzo, J. P., Li, G., & Pollitt, A. (2015). Same data, different perspectives: What is at stake? Response to Savin-Williams and Joyner (2014a). *Archives of Sexual Behavior*, *44*, 15–19.

Li, G., Katz-Wise, S. L., & Calzo, J. P. (2014). The unjustified doubt of Add Health studies on the health disparities of non-heterosexual adolescents: Comment on Savin-Williams and Joyner (2014). *Archives of Sexual Behavior*, *43*, 1023–1026.

Love, B. L. (2017). A ratchet lens: Black queer youth, agency, hip hop, and the black ratchet imagination. *Educational Researcher*, *46*, 539–547.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304–316.

Mayo, C. (2017). Queer and trans youth, relational subjectivity, and uncertain possibilities: Challenging research in complicated contexts. *Educational Researcher*, *46*, 530–538.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*, 403–425.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437–455.

Mittleman, J. (2018). Sexual orientation and school discipline: New evidence from a population-based sample. *Educational Researcher*, *47*, 181–190. doi:10.3102/0013189X17753123

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.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aaac4716.

Raifman, J., Moscoe, E., Austin, S. B., & McConnell, M. (2017). Difference-in-differences analysis of the association between state same-sex marriage policies and adolescent suicide attempts. *JAMA Pediatrics*, *171*, 350–356.

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.

Robinson-Cimpian, J. P. (2014). Inaccurate estimation of disparities due to mischievous responders: Several suggestions to assess conclusions. *Educational Researcher*, *43*, 171–185.

Rosenblatt, J., & Furlong, M. (1997). Assessing the reliability and validity of student self-reports of campus violence. *Journal of Youth and Adolescence*, *26*, 187–202.

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.

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*, 345.e1–345.e15.

Savin-Williams, R. C., & Joyner, K. (2014a). The dubious assessment of gay, lesbian, and bisexual adolescents of Add Health. *Archives of Sexual Behavior*, *43*, 413–422.

Savin-Williams, R. C., & Joyner, K. (2014b). The politicization of gay youth health: Response to Li, Katz-Wise, and Calzo (2014). *Archives of Sexual Behavior*, *43*, 1027–1030.

Schmelling, K. (1995). Averaging correlated data. *Physica Scripta*, *51*, 676. doi:10.1088/0031-8949/51/6/002

Shukla, K., & Konold, T. (2018). A two-step latent profile method for identifying invalid respondents in self-reported survey data. *Journal of Experimental Education*, *86*, 473–488. doi:10.1080/00220973.2017.1315713

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychology Bulletin*, *133*, 859–883.

Vagi, K. J., Olsen, E. O. M., Basile, K. C., & Vivolo-Kantor, A. M. (2015). Teen dating violence (physical and sexual) among US high school students: Findings from the 2013 National Youth Risk Behavior Survey. *JAMA Pediatrics*, *169*, 474–482.

Wimberly, G. L. (2015a). Conclusions and recommendations for further research. In G. L. Wimberly (Ed.), *LGBTQ issues in education: Advancing a research agenda* (pp. 237–251). Washington, DC: American Educational Research Association.

Wimberly, G. L. (Ed.). (2015b). *LGBTQ issues in education: Advancing a research agenda*. Washington, DC: American Educational Research Association.

Wimberly, G. L., & Battle, J. (2015). Challenges to doing research on LGBTQ issues in education and important research needs. In G. L. Wimberly (Ed.), *LGBTQ issues in education: Advancing a research agenda* (pp. 219–235). Washington, DC: American Educational Research Association.

Zaza, S., Kann, L., & Barrios, L. C. (2016). Lesbian, gay, and bisexual adolescents: Population estimate and prevalence of health behaviors. *JAMA Journal of the American Medical Association*, *316*, 2355–2356.

Authors

JOSEPH R. CIMPIAN, PhD, is an associate professor of economics and education policy at the New York University Steinhardt School of Culture, Education, and Human Development, Kimball Hall, 2nd floor, New York, NY 10003; joseph.cimpian@nyu.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.

JENNIFER D. TIMMER, PhD, is an Institute of Education Sciences postdoctoral research fellow in the Department of Leadership, Policy, and Organizations at Vanderbilt University, Peabody College, PMB 414, 230 Appleton Place, Nashville, TN 37203; jennifer.timmer@vanderbilt.edu. Her research focuses on identifying and addressing inequities in education using quantitative methods.