The Effects of Microsuppression on State Education Data Quality

Jacob M. Schauer
Feinberg School of Medicine, Northwestern University

Arend M. Kuyper
Department of Statistics, Northwestern University

E. C. Hedberg
National Opinion Research Center

Larry V. Hedges
Department of Statistics, Northwestern University

Accepted for Publication in Journal of Research on Educational Effectiveness (2020)

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B140042, R305D140045 and National Science Foundation DGE-1437953 to Northwestern University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Introduction

Statewide longitudinal data systems (SLDS) collect an ever-increasing amount of data on their staff and students. A critical function of SLDS is to protect the privacy of individuals, which is not only a guiding principle in many scientific disciplines but is also enshrined in laws such as Family Educational Rights and Privacy Act (FERPA) or Protection of Pupil Rights Amendment (PPRA). To ensure privacy, access to student-level data is often restricted. However, it is understood that such data can be leveraged by researchers to inform improvement and innovation in schools and school systems (Honig & Coburn, 2008; Conaway, et. al., 2015). This presents something of a tradeoff for state education agencies between students' right to privacy and a data-driven approach to enhancing their education. Releasing data to researchers can increase the risks to student privacy, but it is often a crucial step in developing and evaluating educational policies and interventions.

In an attempt to balance these two considerations, states have frequently turned to a measure to mask data called microsuppression (often referred to as "small cell suppression"), wherein certain records are removed from the data prior to its release to external researchers (Seastrom, 2010; Levesque, et. al., 2015). The logic of this approach can be understood in the context of work by Duncan and Lambert (1986, 1989), which frames the risk of disclosing data in terms of an intruder attempting to match released data to target records (for further discussion, see Reiter, 2005). For instance, if an intruder knows that a student was a black female who attended a specific school in fourth grade, then they could query the released data to find black, female fourth-graders in that school. If there is only one such student, then the intruder will be able to know the rest of that student's record, including any sensitive information (e.g., test

scores or special education designations). If there are several such students, then the intruder will be less certain about which is the target record.

Thus, when releasing data, states often divide students into risk strata based on demographic variables that tend to include their school, grade, race, and gender. Strata with few students are dropped from the data prior to its release. Typically, “few” will mean fewer than 10 (Levesque, et. al., 2015). This means that if an intruder matches a student’s school, grade, race, and gender in the released data, then there will be at least 10 records that match, and so the intruder will be less certain about which one pertains to the target record they are seeking.

This masking method is applied to both raw data and summary statistics. For example, tables published in journals or on websites of state departments of education may not report values in cells with small counts or extreme values (e.g., Delaware, 2020; North Carolina, 2009). To protect public-use datasets available through state departments of education, which can contain student-level data, states suppress records for students in small cells (e.g., Massachusetts, 2014). We have even engaged in projects where suppression rules have been applied to student-level data shared under secure data use agreements (DUA). In other words, state data administrations can and do apply microsuppression procedures to many of the data types available to researchers.

While microsuppression has seemingly satisfied state concerns over disclosure risks, less is known about how this affects the utility of the data it produces. It is entirely possible that by deleting certain records, the released data no longer resembles the original data in important ways. This could potentially lead to analyses or inferences based on microsuppressed data that differ from those conducted on the complete (non-suppressed) data. However, there is little

empirical evidence about how much these may differ in US state education data, and what may be driving those differences.

This paper addresses how microsuppression can affect the quality of data in education. The following section provides an overview of microsuppression and how states have employed it. We then discuss how microsuppression can create "biased" data and argue that biases depend on how much and which data gets suppressed. Using cross sectional data from states we examine potential biases in various statistics computed from microsuppressed data. We find that some marginal summary statistics, such as mean test scores for the entire state, exhibit minimal bias, but that conditional estimates, such as mean test scores for minorities, can exhibit substantial bias. Moreover, we demonstrate potential corrections, such as post-stratification, may actually do more harm than good that in some instances. Finally, we discuss the implications of these findings and alternative measures that SLDS may take.

Microsuppression for Data Masking

Data privacy is a central focus of many efforts that collect data on individuals, especially scientific studies and government entities like the US Census Bureau. Ensuring an individual's right to privacy is often treated as both a moral and legal obligation. It is also understood that doing so can help ensure data collected are of high quality because respondents trust that their identity and responses are protected (see Hundepool et al., 2012). SLDS and other educational agencies view privacy in similar terms, and students' privacy is legally protected by FERPA and PPRA. In addition, Institutional Review Boards (IRBs) also evaluate and govern how potential research efforts must protect individual privacy.

The right to privacy is particularly important given the vast amounts of sensitive information collected by state agencies on students. In addition to students' progress and achievement throughout their primary and secondary education, states and schools often track how students access academic or counselling services. These records alone present "big data" problems when it comes to storage and security (Trainor, 2015). An increasing number of states, such as Illinois, Minnesota, or Maryland, capture (and can link education data to) earnings and employment status later in life (see, e.g., Illinois Longitudinal Data System). Moreover, with the widespread proliferation of technology-based tools in education, the amount of data collected on any one student has grown by orders of magnitude. In short, SLDS and other education stakeholders are storing and maintaining vast amounts of data on students, and they have a serious obligation to maintain the privacy of that data.

In order to live up to this obligation, SLDS have taken important steps. They store data on secure servers and restrict access to that data. However, states are seldom immune from *ever* granting access to students' data. Public reporting, including for accountability purposes, is a common use of SLDS data. Researchers and state agencies often use SLDS data in order to evaluate various educational policy changes. Thus, SLDS enact sophisticated measures to ensure that when data are accessed and reported, unintended disclosures are unlikely. These measures include security protocols for transferring data to researchers, or even requiring that researchers conduct analyses in specific physical locations that are deemed secure, which is the approach favored by Arizona (see Arizona, 2020). When data is shared with external researchers, states typically remove personally identifiable information, such as names and addresses, and evaluate the risk that the remaining variables may be combined to identify individual students (Johnson,

2007). This process has even been built directly into software that queries raw data, such as Delaware's systematic redaction system (see Peoples, 2018).

As part of this effort, SLDS have turned to a data masking tool called microsuppression to protect against disclosures. Microsuppression, which was introduced by Duncan and Lambert (1986), is a procedure that removes records at high risk of disclosure from a dataset prior to its release (e.g., to external researchers). One way to understand the disclosure risks is to consider an "intruder" seeking sensitive information on a specific student or set of students (Duncan & Lambert, 1986, 1989). The intruder may know some information about that student, such as their grade, gender, etc., which can be denoted by a vector $\boldsymbol{t}$ (see Reiter, 2005). The vector $\boldsymbol{t}$ may not contain sensitive information about the target student, but it can be used to uncover sensitive information by matching $\boldsymbol{t}$ to records in the released data. If only one or two records match the information, the intruder has a high probability of uncovering the rest of the student's record, which can include sensitive information. However, if $\boldsymbol{t}$ matches 20 records, then the intruder has a considerably lower probability of uncovering the rest of the student's record; it could be any of the 20 records matched. So long as the values of any sensitive variables in those 20 records are not identical, this presents a reasonable approach to protecting privacy, and the risk of an intruder uncovering sensitive information will be greater when they can identify very few, or even a single record that match their information $\boldsymbol{t}$.

To protect against this, microsuppression works by deleting records for which a potential intruder's vector of information $\boldsymbol{t}$ has only a few matches. This is related to an idea called *k*-anonymity in the statistical disclosure control literature (see Samarati, 2001; Samarati & Sweeney, 1998). How this occurs, and how it can affect the utility of the resulting data will depend on what variables are used to divide students into risk strata (i.e., cells), and what size of

strata is considered small enough to delete. Because of the diversity of state education systems and the data requests they receive, SLDS vary in how and when they use microsuppression. However, there appear to be at least two different settings in which SLDS apply it.

First, microsuppression is applied is to student-level data, and occasionally to data shared with external researchers under secure DUAs. Based on a review of data governance manuals, correspondence with SLDS administrators, and experience in our own research, it appears that microsuppression of restricted access student-level data either has occurred or may occur in at least five states (and potentially even more than five) that comprise nearly 20% of all students enrolled in public schools in 2020.

Second, and more common, microsuppression is applied to summary tables. SLDS produce a wide range of publicly available tabular data that summarize everything from student achievement and demographics to school staffing. Many of these tables include school, district, or county averages (including for subgroups) and many of these are publicly available. For instance, North Carolina produces "School Performance Grades Reports" (see North Carolina, 2020) that contain average student scores on state assessments by demographic subgroups within schools. These tables show, for instance, how the average English language learner in each school performed on the state science exam. When these tables are publicly available, they are almost always subjected to some type of suppression. Every SLDS administrator we contacted and every data governance manual we reviewed, totaling over 25 states, contains some form of suppression rules for publicly reported tables.

The exact procedure used for this are different from state to state. For instance, Delaware uses a set of rules that apply both to cell counts in tables as well as cell counts in the population. Under these rules, cells in tables of fewer than five students will be redacted. However, even if a

cell contains more than five students, if that cell is representative of a population subgroup with fewer than 15 students, it will automatically be redacted (Delaware, 2020). A more common rule appears to be that cells with fewer than 10 individuals are suppressed, which is in data governance manuals for several states (e.g., Massachusetts, 2014; Montana, 2018; Nebraska, 2013; North Carolina, 2009). As an example, if there are only one or two ELL students in a school, North Carolina will not report the average test scores of ELL students in that school in their "School Performance Grades Reports."

Microsuppression and Data Quality

Various researchers in different fields have pointed out that microsuppression can degrade the quality of the released data (Kelly, 1992; Ohno-Machado, 2002; Matthews, et. al., 2017). One way to conceive of the quality of microsuppressed data is to examine the ways in which they differ from the complete, unsuppressed data. If there are large differences between the complete and microsuppressed data, this can lead to inaccuracies in the results of analyses conducted on the microsuppressed data, which could affect scientific or policy conclusions supported by those analyses. Thus, it is important to know if and when the released data can be considered a reasonable proxy for the complete data.

To help understand issues surrounding data utility, it can be helpful to think about microsuppression involving three different subsets of the data. There is the *complete* dataset that contains, for instance, entries for every student in the state. Then there is the dataset of dropped observations, which correspond to students whose records fall into small strata and thus are suppressed prior to release. Then there is the data that ultimately gets released. Throughout this paper, we will refer to these as the complete/full, dropped, and released/retained data. The

dropped and released data are mutually exclusive (i.e., no record is both released and dropped), and their union is just the complete dataset.

In this article, we compare the released and complete data by examining population-level statistics for each state. If key statewide statistics are not preserved in the released data, then we might think of the released data as "biased" in some way. The following sections detail some potential quantities of interest, how microsuppression might affect those parameters, and the statistical factors that could induce biases.

Bias

We want statistics computed from state datasets to be unbiased. One can conceive of bias in analyses of released data in terms of three parameters: μ_r in the released dataset, μ_d in the dropped dataset, and the population parameter μ in the complete dataset. To get a sense of this, suppose we are interested in the average math achievement test score for minorities in fourth grade in a given state. Let Y be the math scores in the state, and let X indicate whether a student is a minority ($X = 1$ corresponds to minority status). Then the estimand we are interested in is

$$\mu = E[Y \mid X = 1] \tag{1}$$

If we had the full dataset, we might compute this with the conditional mean of test scores for minorities. This would give us the value of μ. However, if suppression has occurred, we do not have access to the complete dataset. Let S be an indicator for whether or not an individual is in a small stratum. Then the released data would comprise only observations for which $S = 0$. Thus, if we were to take the average test score for minorities in the released data, we would have

$$\underline{\mu_r} = E[Y \mid X = 1, S = 0] \tag{2}$$

Additionally, we can denote the mean test score for minorities in the dropped data (i.e., among the records that are suppressed prior to release) as:

$$\mu_d = E[Y \mid X = 1,\ S = 1] \tag{3}$$

The mean in the complete data can be different from that in the masked data, and thus it may be difficult to use the masked data to make inferences about μ. Indeed, we can express μ as a function of μ_r and μ_d using the law of total expectation:

$$\mu = \mu_r\, P[S = 0 \mid X = 1] + \mu_d\, P[S = 1 \mid X = 1] \tag{4}$$

The difference between μ and μ_r can be referred to as bias. Moreover, we can rewrite equation (4) to obtain a formula for the bias:

$$\text{Bias} = (\mu_r - \mu_d)\, P[S = 1 \mid X = 1] \tag{5}$$

This expression contains two components. The first is the difference between the released ($S = 0$) and dropped ($S = 1$) test scores $\mu_r - \mu_d$. If the retained and suppressed records have the same mean test score, then the bias will be zero. If, however, they have different means, then there may be systematic differences between the released and dropped observations. This can result in bias, but it also means that correcting for this bias may be difficult. This is because the observations that are dropped from the dataset prior to its release may differ systematically, and in ways unknown to the researcher, relative to the observations that are released.

The second component $P[S = 1 \mid X = 1]$ is the probability that an observation is dropped, what might be called a suppression rate or the proportion of data suppressed (PDS). The suppression rate is an important quantity here. If it is 0%, then the bias is zero since no observations are dropped. If it is larger, then there may be substantial bias.[1]

[1] Note that if the suppression rate is 100%, which occurs for some racial subgroups in the data, then an analysis involving those subgroups will be impossible.

Variance

Another type of parameter that may be of interest to researchers is the variation in the data, and how this variation is partitioned at different organizational levels. Typically in education, we focus on the total variation and the intra-class correlation (ICC), which refers to the correlation between observations in the same school or classroom.

These parameters are important for a few reasons. Both the total variation and the ICC are used to justify the design of experiments in education. If such values are computed on released data, then they may not correspond to the "true" values from the complete data. Inaccuracies in these parameters could lead to experiments that are either under- or overpowered. Alternatively, analyses involving a stochastic element (e.g., with sub-samples of the data or a probability model) must account for some level of uncertainty, which is usually communicated as a standard error. The standard error of an estimate will often depend on the total variation, the ICC, or both. Thus, if such analyses are conducted on the released data, the standard errors may be incorrect, since there is no guarantee that either parameters will be unaffected by microsuppression. This will in turn affect the properties of hypothesis tests; if the ICC in the released data is much smaller than the ICC in the complete data, tests may have larger than nominal type I error rates.

In this study, we consider school-level ICCs. Denote the variation of average test scores between schools as τ^2 and the variation among students within schools as σ^2. Then the total variation can be written as $\tau^2 + \sigma^2$ and the ICC can be expressed as

$$\rho = \frac{\tau^2}{\tau^2+\sigma^2} \tag{6}$$

In the released data, the corresponding parameters are τ^2_r and σ^2_r, and the ICC in the masked data can be written as

$$\rho_r = \frac{\tau_r^2}{\tau_r^2+\sigma_r^2} \quad (7)$$

Suppression can affect the value of both τ^2 and σ^2, and thus ρ. For example, value of σ^2 may decrease if students whose records are dropped are in the tails of the achievement distribution, but it can increase if they are closer to the center of that distribution. The between-school variation can be affected by the deletion of individual students within schools if those deletions render school means more similar. Moreover, it may be the case with smaller schools that all of the students in the school (and hence the school itself) are dropped from the data due to suppression. If the dropped schools are particularly high (or low) achieving, then this may decrease τ^2 (and hence ρ).

Data & Methods

This article empirically assesses the potential impact of microsuppression using state education data. Our approach is to mimic the behavior of microsuppression procedures using that data. We begin with complete data. The for each dataset, we divide records into risk strata and suppress the small strata according to rules used by states when releasing data. We then run analyses on the masked data and the complete data and compare their results. This section details the data used, the procedures for suppressing data, and some limitations of the data and our approach to studying microsuppression.

We obtained data from eight states. For each of these states, we took cross sections of fourth graders and eighth graders. Since the data from each state covered slightly different time frames, the years of the cross sections vary for each state; the resulting cross sections span the

years from 2009 to 2012 and include elementary and middle school students. These cross sections contain basic demographic information, including a student's school, race, gender, if they receive free or reduced-priced lunch (FRL), and whether they have limited English proficiency (LEP). They also include state achievement test scores in various subjects, though all datasets (across states and grades) have scores for math and reading. Table 1 provides a summary of cross sections used.

Microsuppression was conducted by stratifying students based on their school, race, and gender. We then deleted records in strata smaller than 10. Deleting strata with fewer than 10 students follows the guidance of the privacy and data management plans of the states that shared data for this effort and is consistent with the policies of many other states as described in the previous section.

Two limitations are worth noting about this approach. First, the results presented here are based on only eight states. While they range from the very small (State 5) to large (State 4), and from racially diverse (State 1 and State 6) to homogenous (State 5 and State 8), it is somewhat unclear how the results presented in this article generalize to other states. Second, while our approach to microsuppression is consistent with practices in the states we studied, it is not the only way to apply suppression rules. It is possible to, for instance, include additional demographic variables, such as ELL status when defining risk strata, or to include suppression rules that involve extreme values. Thus, the results presented in the subsequent sections show *possible* impacts of microsuppression on data utility, and these impacts might vary had we used data from different states or applied different suppression procedures.

Empirical Results

The sections that follow provide insight into the effects of microsuppression. We compare marginal and conditional means, as well as variance components of the released data to those of the complete data. To get a better sense of why differences emerge between released and complete datasets, we examine suppression rates, as well as systematic differences between the dropped and released data. Note that because each cross section is a census of students in a state (for a given grade) quantities in the following sections are reported without standard errors. However, an alternative conception of these data would treat it as a random sample. Had we done so, standard errors would largely be less than about 0.01 for marginal statistics and 0.05 for subgroup statistics.

Mean Test Scores

A basic question about microsuppression is how well population-level statistics are preserved. In the data, there are both continuous variables (test scores) and categorical variables. For test scores, we compute the difference between average test scores from the masked and complete data on the scale of standard deviation units:

$$\delta = \frac{\mu_r - \mu}{\sigma} \tag{8}$$

where μ_r and μ are the mean scores for the released and complete data, respectively, and σ is the standard deviation of the test scores computed on the complete data. For categorical data, we present the raw proportions for each subgroup in the population, such as the proportion of students receiving FRL.

Figure 1 shows the differences in test score means for each cross section of data in a series of plots. For both the fourth and eighth grade cross sections, Figure 1 shows the suppression rate, the difference between the released and dropped test scores for math and

reading, and the difference between the released and full data test scores. Differences are reported in standard deviation units. In the plots that show the difference between the released and full data (last two plots in each row), we see that while the average test scores in the masked data are all greater than those of the complete data (i.e., all of the differences are positive), the differences appear quite small in magnitude. The largest differences are just under than 0.05 standard deviations (State 7, fourth grade), and most are below 0.02 standard deviations.

These (apparently minor) differences will depend on both the suppression rate and the differences between the records that are released versus those that are dropped. Figure 1 provides insight into how these relate to state education data. The plot of suppression rates (first plot in each row) show that they vary greatly across states. Some states are required to delete as much as a quarter of their data under the suppression rules, while others suppress only 5% of it. There is no single driving factor for these differences. For instance, State 5, which has high rates of suppression, is a small state with less racial diversity (see Figure 1) and fewer than 25 students per school (in a single grade) on average. When fewer students are in a school, dividing those students into strata will often create small strata that will need to be deleted. Conversely, States 1 and 6, both of which also had high rates of suppression, are larger, more diverse states, and schools in those states had more than 60 students (per grade) on average.

To further unpack what type of student records get suppressed, we ran multilevel probit models to predict whether a student's record was suppressed based on their characteristics and the characteristics of their school. These models, which included random intercepts for schools, revealed that, as argued above, school-level characteristics, including mean achievement scores for schools, are often more predictive of suppression than individual characteristics. In general, these models suggested two patterns. In some states, lower achieving students in higher

achieving schools were more likely to be suppressed. In other states, lower achieving students in lower achieving schools were more likely to be suppressed. In other words, while students with lower test scores (for their school) were more often suppressed, in some states these students more frequently attended higher achieving schools, and in other states these students more often came from lower achieving schools. Because race and socioeconomic status are correlated with achievement, we tended to find greater discrepancies between the complete and released data for students in certain subgroups, which we discuss in subsequent sections.

An important related issue is that the differences between the released and dropped records ("Released – Dropped" plots) show that the records retained can exhibit substantially higher test scores than those that are dropped. For several cross sections, these differences are as large as 0.3 standard deviations in magnitude. For reference, 0.3 would be considered a small-to-medium sized effect in the social sciences (Cohen, 1988), and comprises approximately a year's worth of learning in math in grade 4 (Hill, et. al., 2007).

While both the suppression rates and the Released-Dropped differences are each related to the Released-Full differences, they are not related to each other in the data. Viewing the suppression rate and "Released – Dropped" plots together, we see that there is only a weak correlation between them (r = -0.06 for math and 0.03 for reading). Indeed, some states suppress modest amounts of data that differ in large ways from the data that gets retained (e.g., State 7, eighth grade), while for other states who delete about the same proportion of data their deleted records are quite similar to the ones they retain (e.g., State 3).

Variance Components

Both the planning of experiments and the precision of certain analyses will depend on the variation in the data. In this paper, we examine the effect of microsuppression on the total variation, as well as on the ICC. Figure 2 compares the differences of these two quantities for test scores in the cross sections of data used in this article. The left plot in Figure 2 contains a dot that corresponds to the percent difference in the total variation between the released and full data for each grade, state, and test. Negative values indicate that the released data exhibit less total variation than the complete data. The figure shows that the released data frequently have less variation than the complete data. While for some of the datasets this reduction is modest (less than 2%), this is not universally true, as microsuppression reduced the total variation of test scores for eighth graders in State 2 by as much as 11.5%.

The right plot in Figure 2 shows results for ICCs. This plot contains dots that reflect the percent difference in ICC between the released and complete data for each state, test, and grade. For fourth grade test scores, there are substantial differences in ICCs, as large as 31% in magnitude. However, while for some states and tests this difference in positive, meaning that the released data have larger ICCs, for others it is negative, meaning that the ICCs are smaller in the released data. This contrasts with the results for eighth graders, where the released data almost universally have smaller ICCs, and in some instances substantially so.

Bias in the variance components can affect many of the statistical procedures commonly used in education research. For instance, Hedges (2007) and Korendijk et al. (2010) show that when assessing treatment effects in cluster randomized trials, an inaccurate ICC can substantially inflate the type I error rate of the resulting hypothesis test. Similarly, if the total variation is systematically underestimated, it can induce overestimates in the precision of effect estimates or elevate type I error rates of hypothesis tests.

To put the findings in Figure 2 in context, consider tests of treatment effects in cluster randomized trials. Based on the results of Hedges (2007), underestimates of variance components on the scale of State 2 (30% decrease in the ICC and 11% decrease in the total variation) can raise the type I error rate (i.e., false positive rate) of such tests from the nominal 5% to between 10% and 15%, depending on the number of schools and students in an experiment. Conversely, when the variance components are overestimated as much as they are among State 4 fourth graders (21% increase in ICC, 1% increase in total variation), the test for treatment effects will be more conservative (type I error rate as low as 4%) but will also be considerably less powerful (see Guittet, et al., 2005).

Demographics

Released data contains fewer FRL and LEP students, as well as a smaller percentage of minorities when compared to the complete data. Figure 3 shows the differences between datasets for the proportion of LEP, FRL, or female students for a given state and grade. The left plot shows the difference between the released and complete data. Each dot in that plot corresponds to a given cross section of data and demographic variable, and differences are computed as raw differences in percentages. The differences between the composition of the full and released data for these variables tends modest (i.e., less than 2%). However, the right plot in Figure 3 reveals systematic differences between the released and dropped records. Each dot in the bottom plot shows the difference between the released and dropped records computed on each cross section of data and reported as a difference in percentages. The distribution of those differences shows that students whose records get suppressed tend to be much more likely to receive FRL or be designated LEP than students whose records are released.

More sizable differences occur for race. Figure 4 shows the proportion of each race present in each cross section of complete ("Full") and released data ("Released"). The figure shows that released data tends to contain more white students and fewer minority students. The difference between the racial composition of the full and released data can be substantial. In State 6, the released data on fourth graders would contains nearly 66% white students, while the full data has only about 50% white students. In State 1, the released data on fourth graders would contain nearly 2% black students and less than 1% Asian students compared to roughly 8% and 5%, respectively, in the full data. The percentage of Native American students in the released data is typically less than half of that in the full data. Some racial categories are completely suppressed (i.e., no such records are released), such as with Hispanic students State 5 or Native American students in State 8.

Finally, it is worth noting that subgroups defined by multiple demographic variables appear in released data at frequencies not too different from the complete data. For instance, the proportion of minority students receiving FRL in the released data is typically 2-3% lower than that of the complete data. The same can be said of LEP status.

Conditional Means

Microsuppression can also impact conditional distributions, including the distribution of test scores for subgroups in the data, such as race, gender, or FRL status. Suppose an analysis involves the mean test score for students receiving FRL. If the conditional mean test score for students receiving FRL in the full data $E[Y \mid FRL]$ differs greatly from that in the released data $E[Y \mid FRL, S = 0]$, this can greatly affect the accuracy of such analyses.

Figures 5 and 6 show how mean test scores for each racial subgroup differ across datasets. The top panel of each figure plots the standardized mean difference between the released versus complete data computed within each racial subgroup for each state, grade, and test. Each bar corresponds to a given state, grade, and race. Bars above zero indicate that the conditional mean for the released data exceeds the conditional mean in the complete data, and bars below zero indicate that the released data has a smaller mean than the complete data. Both figures show that the test scores of white and Asian students are often overestimated by the released data, but the test scores of black, Hispanic, and Native American students are underestimated. These differences can be large in magnitude. The difference in mean test scores between the complete and released data is on the order of about 0.1 to 0.2 standard deviation units for black and Hispanic students and can be as large as 0.35 for Native American subgroups.

These differences are driven both by the suppression rates for these subgroups and differences in the released versus dropped records. Recall from the previous section that a greater fraction of black, Hispanic, and Native American student records are dropped during microsuppression, so suppression rates will be high for these subgroups. Moreover, the lower panels of Figures 5 and 6 show the difference in average achievement between individuals whose records are released versus those who are dropped for each race. What can be seen in these panels is that for most states and grades, the minority students whose records are released score on average about 0.2 standard deviations lower than those who are suppressed. Meanwhile, the white and Asian students whose data are released tend to score 0.2–0.5 standard deviations higher than those whose data are dropped. In fact, for one state, test scores released on white students are on average almost a full standard deviation higher than the test scores for students

whose data are suppressed. This would seem to be consistent with the relationship between suppression and school context and individual characteristics discussed in previous sections.

Taken together, the released data tend to be missing many lower achieving white students and higher achieving black and Hispanic students. This widens the apparent racial achievement gaps in the released data relative to the complete data. Figure 7 show the differences for each state and grade for the white-black (left plot) and white-Hispanic (middle plot) gaps. Each plot is comprised of dots that reflect the standardized difference achievement gaps between the released and complete data for a given cross section. Differences are reported on the scale of standard deviation units. Positive values in these plots correspond to released datasets that overstate the racial achievement gap. We see that for most states, this overstatement is modest (less than 0.05), but it can be as large as 0.2.

For the other demographic variables, differences in achievement gaps between complete and released data range from small or large. For instance, the difference in mean test score for students who do not receive FRL versus those who do will be overstated in the released data, but only by about 0.01 standard deviations. However, the achievement gap between non-LEP and LEP students may be substantially different. The right plot of Figure 7 shows the differences in the non LEP-LEP achievement gaps in the released and complete data; values are computed on the scale of standardized mean differences as in equation (8). Note that for math scores, the released data will often overstate the achievement gap (i.e., the difference between the complete and released values is negative) by as much as 0.1 to 0.2 standard deviations. For reading, these differences are considerably more modest. This is due in part to the fact that many LEP students are missing reading scores in the data (possibly due to not taking the reading achievement tests),

including among deleted observations. For both math and reading, these differences are larger in magnitude in contexts where a greater proportion of LEP students data get suppressed.

Post-Stratification Weighting

The previous sections detail the ways in which data subjected to microsuppression can be biased. This is due to the fact that records that are released after microsuppression are not representative of the complete dataset. This is analogous to issues with survey sampling, where the composition of a sample may differ from that of a population. A common correction for this problem involves post-stratification weighting so that the sample (i.e., the released data) more closely resemble the population (i.e., the complete data). This involves dividing the data into post-hoc strata and weighting observations within strata proportional to the stratum size (see Lohr, 2006; Gelman & Carlin, 2000). A similar approach could be taken with microsuppressed data; for instance, if after suppression, there are fewer Hispanic female students in the data, then we can upweight observations that correspond to Hispanic females. Note here that the strata used to weight the data are not necessarily the same as those used to mask or suppress data.

Forming post-stratification weights would be difficult for at least two reasons. First, determining the weights requires some knowledge about population-level counts. As an example, figuring out how much to upweight records from Hispanic female students involves knowing how many Hispanic females are in the complete data. This type of information inherently missing from data subject to microsuppression, and so requires information beyond merely the data released. Some of it can be obtained in administrative reports, for instance from the National Center for Education Statistics' Common Core of Data (CCD).

The second difficulty concerns how strata are formed for the post-stratification weights. They should be granular enough to minimize differences between the released and complete data. However, if they are too granular, then there will be strata in the population for which there are no corresponding observations in the released data. For instance, if we felt that race and gender were sufficient to determine post-stratification weights, but all of the Hispanic females are deleted from the released data due to microsuppression, then there will be no observations to upweight for that stratum.

A first-order approach to applying these weights could be to use student-level demographics, such as their race, gender, and school they attend. Weights can be computed from information in CCD reports on demographics within states. In this section, we demonstrate how this approach might proceed, and what improvements it might offer. Let $\boldsymbol{Z}$ be the set of variables in the data that are used to divide it into strata for post-stratification weights. In this article, $\mathbf{Z}$ will comprise either a student's race and gender ($\boldsymbol{Z}_1$), their school ($\boldsymbol{Z}_2$), or all three ($\boldsymbol{Z}_3$). Then, from the full data, we can determine the proportion of students with the same set of covariates $\boldsymbol{z}$ as $P[\boldsymbol{Z} = \boldsymbol{z}]$, and reweight observations in the masked data by $1/P[\boldsymbol{Z} = \boldsymbol{z}]$. We follow this procedure to obtain corrected mean test scores for each state and grade.

Differences between weighted mean test scores in the masked data and the mean test score in the complete data are shown in Figure 8. For each test, the figure shows the bias of various weighted corrections denoted by dots, which are shaded by the variables used to compute post-stratification weights. Corrections include an unweighted mean (the "None" dots), which is how Figure 2 is computed, and a mean that weights observations by race and gender; by school; and by race, gender, and school ("All"). Values are reported in terms of standardized mean

differences as computed by equation (8); positive values indicate that the masked-data mean is larger than the complete-data mean.

What can be seen in the figure is that incorporating race and gender into the weights can actually exacerbate bias. The differences in those columns are larger than those in the unweighted column, meaning that less bias is obtained by not weighting by race or gender. For some states and grades, means weighted by race or gender are as much as 0.3 standard deviations larger than the complete data mean. This bias can be reduced by weighting only by school size. However, even there, the unweighted mean tends to be slightly less biased. Weighting by all three increases the bias relative to weighting just by school size.

The increase in bias arises from the fact that the dropped and released records systematically differ. A weighting correction works by giving more weight to certain observations under the assumption that those observations are representative of several records that are *not* observed. But much of this article has shown that this assumption does not always hold with microsuppression. In Figures 5 and 6, we saw that there were substantial differences (on average) for racial subgroups between the released and dropped observations, and these differences were particularly large for black and Hispanic students. At the same time Figure 1 shows that a greater proportion of black and Hispanic students are removed from the data under the suppression rules. Thus, the black and Hispanic students whose records are released will receive greater weight under the correction, but they will have higher test scores than black and Hispanic students whose records are dropped. In that sense, the correction gives *more* weight to observations that are arguably *less* representative.

Discussion

Protecting individuals' privacy is an important principle in any effort that systematically collects and uses data. This is part of the reason that laws such as FERPA and PPRA, as well as bodies like IRBs require that researchers and state agencies take measures to protect students' privacy. Many of these measures are built into the SLDS infrastructure and are key steps in how they provide data to the public and to researchers. One such measure, which appears to be common practice, is microsuppression. It has long been known that microsuppression can limit the utility of data, however the extent to which it can has not been fully studied in the context of SLDS.

This paper has sought to quantify the effect of masking state education data via microsuppression. While the data and suppression rules used in this study do not cover all possible scenarios, they do offer some idea of the implications of microsuppression on data quality for SLDS. We found that the data that results from microsuppression are often biased, and we would urge caution when interpreting analyses conducted on such data. In particular, we found that while microsuppression had almost no impact on marginal mean test scores, it did substantially reduce measures of variance. Underestimates in both the total variation and the ICC found in this article would be enough to distort the results of analyses such as null hypothesis tests in cluster randomized trials. Microsuppresion also induced greater biases within subgroups. The achievement of white and Asian students, for instance, tends to be higher in the released data, while the achievement of black and Hispanic students tends to be lower. This can lead to sizeable overestimates of achievement gaps as large as 0.2 standard deviations.

The key factor driving these biases is that the released and dropped records are systematically different in ways that are unknown to the analyst. Because of this, potential corrections based on post-stratification tended to exacerbate biases rather than mitigate them.

This implies that not only might analyses based on microsuppressed data be biased, but there will be no obvious way to adjust for that bias.

While microsuppression has found favor among SLDS, it is only one tool in a broader literature on statistical disclosure control. There are numerous alternative measures that SLDS could take to mask data (for a larger discussion, see Hundepool et al., 2012). For instance, agencies could release averages of randomly grouped individuals (referred to as "microaggregation"), which can preserve data utility better than microsuppression (see Card et al., 2010; Matthews & Harrell, 2011), including for analyzing educational experiments (Schochet, 2020). Synthetic data methods have also emerged as a potentially useful alternative. These methods generate sets of synthetic data by randomly imputing values for, say, a student's test scores or demographics (see Little, Liu, and Raghunathan, 2004; Raghunanthan, Reiter, & Rubin, 2003; Reiter, 2004, 2005b, 2009; Reiter & Raghunathan, 2007; Rubin, 1983; Singh, Yu, & Dunteman, 2003). A recent empirical evaluation of synthetic data disclosure conducted on student records in Maryland highlights the promise of this approach and discusses steps required to ensure successful implementation (Bonnéry et al., 2019).

State agencies may also address issues of data utility using a system of verification. Because the biases induced by microsuppression are seldom reported by states themselves, it will be difficult to say just how accurate a given analysis conducted on microsuppressed data actually is. However, states could potentially run the same analysis on the complete data and quantify those biases post-hoc. Related work by Reiter et al. (2009) and Reiter (2018) discuss verification systems wherein synthetic data are released, researchers use those data to train models, and then submit their code to data administrators who can verify analytic results on the real data.

Adopting any of these alternatives is not necessarily a trivial process. In some states, like Delaware, the machinery for conducting suppression is built directly into the tools that interface with raw data. Moreover, methods such as synthetic data or verification systems may require additional personnel, expertise, and/or computational power. Thus, while SLDS may have other options in theory, it may take substantial effort to make them common, reproducible practice. In the meantime, the results of this article suggest that transparency about data collection and masking is necessary in order to properly contextualize the results of analyses involving state education data.

References

Arizona Department of Education. (2020). Data management. https://www.azed.gov/data/.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. And Tourangeau, R. (2013). Report of the AAPOR Task Force on Non-probability Sampling. Technical report, American Association for Public Opinion Research, Deerfield, IL.

Bonnéry, D., Feng, Y., Henneberger, A. K., Johnson, T. L., Lachowicz, M., Rose, B. A., Shaw, T., Stapleton, L. M., Woolley, M. E., & Zheng, Y. (2019). The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *Journal of Research on Educational Effectiveness*, *12*(4), 616-647.

Conaway, C., Keesler, V., & Schwartz, N. (2015). What research do state education agencies really need? The promise and limitations of state longitudinal data systems. *Educational Evaluation and Policy Analysis*, *37*(1S), 16S–28S.

Chetty, R., Feldstein, M.S., & Saez, B.E. (2010). Expanding access to administrative data for research in the U.S. Washington, DC: National Science Foundation 10-069 White Paper.

Delaware Department of Education. (2020). Data privacy, redaction. https://www.doe.k12.de.us/Page/3024

Drechsler, J., Bender, S., Rässler, S. (2008). Comparing fully and partially synthetic datasets for statistical disclosure control in the german IAB establishment panel. *Trans. Data Privacy*, 1(3), 105–130

Duncan, G. T. & Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, *81*, 10–28.

Duncan, G. T. & Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, *7*, 207–217.

Elliott, M. R. & Valliant, R. L. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249–264.

Guittet, L., Giraudeau, B., & Ravaud, P. (2005). A priori postulated and real power in cluster randomized trials: mind the gap. *BMC Medical Research Methodology*, *5*(25).

Herzog, T. N. & D. B. Rubin. (1983). Using multiple imputations to handle nonresponse in sample surveys. In *Incomplete Data in Sample Surveys*, edited by W.G. Madow, I. Olkin, & D. B. Rubin, 2:209–45. Academic Press.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2007). Empirical benchmarks for interpreting effect sizes in research. MDRC Working Papers on Research Methodology. Retrieved from https://www.mdrc.org/sites/default/files/full_84.pdf.

Honig, M. & Coburn, C. (2008). Evidence-based decision making in school district central offices. *Educational Policy*, *22*, 578-608.

Illinois Longitudinal Data System. https://www.illinoisworknet.com/ILDS/Pages/default.aspx

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples*, edited by H. Wainer, 115–42. Springer-Verlag.

Johnson, C. (2007). Safeguarding against and responding to the breach of personally identifiable information. Memorandum for the Heads of Executive Agencies. U.S. Office of Management and Budget.

Kaciroti, N. A. & T. E. Raghunathan. (2014). Bayesian sensitivity analysis of incomplete data: Bridging pattern-mixture and selection models. *Statistics in Medicine*, *33*(27), 4841–57.

Kelly, J. P., Golden, B. L. and Assad, A. A. (1992), Cell suppression: Disclosure protection for sensitive tabular data. *Networks*, *22*, 397-417.

Korendijk, E., Moerbeek, M., & Maas, C. (2010). The Robustness of Designs for Trials With Nested Data Against Incorrect Initial Intracluster Correlation Coefficient Estimates. *Journal of Educational and Behavioral Statistics*, *35*(5), 566-585.

Lee, S. & Valliant, R. L. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3), 319–343.

Levesque, K., Fitzgerald, R., & Pfeiffer, J. (2015). A guide to using state longitudinal data for applied research. National Center for Education Evaluation and Regional Assistance Report # NCEE 2015–4013. U.S. Institute for Education Sciences.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88 (421), 125–34.

Little, R. J. A. (2009). Selection and pattern-mixture models. In *Longitudinal Data Analysis*, edited by G. M. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs, 409–31. Boca Raton, FL: CRC Press.

Little, R. J. A., Liu, F., & Raghunathan, T. E. (2005). Statistical disclosure techniques based on multiple imputation. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, edited by W. A. Shewhart, S. S. Wilks, A. Gelman, and X. Meng.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data (2nd edition)*. Hoboken, NJ: Wiley.

Lu, B. & Lemeshow, S. (2018). Survey sampling and propensity score matching. In *The Wiley Handbook of Psychometric Testing*, edited by P. Irwing, T. Booth and D. J. Hughes.

Manski, C. (2007). *Identification for prediction and decision*. Cambridge, Mass.: Harvard University Press.

Massachusetts Department of Elementary and Secondary Education. (2014). Researcher's guide to Massachusetts state education data. Retrieved from http://sites.bu.edu/miccr/files/2015/12/Researcher-Guide-to-Massachusetts-State-Education-Data.pdf.

Matthews, G. J. & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, 5, 1–29.

Matthews, G. J., Harel, O., & Aseltine, R. H. (2017). A review of statistical disclosure control techniques employed by web-based data query systems. *Journal of Public Health Management and Practice*, *23*(4), e1–e4.

Montana Office of Public Instruction. (2018). Montana's consolidated state plan under the Every Student Succeeds Act. Retrieved from http://opi.mt.gov/Portals/182/Page Files/ESSA/Goodbye NCLB,Hello ESSA/Accessible ESSA Submission Jan 2018 Updated Date.pdf.

Nebraska Department of Education. (2013). Data access and use policy and procedures including research and evaluations. Retrieved from https://2x9dwr1yq1he1dw6623gg411-wpengine.netdna-ssl.com/wp-content/uploads/2017/07/Nebraska_Data_Access_and_Use_Policy_and_Procedures.pdf.

North Carolina Department of Education. (2009). Data management group policy: Reporting on data in small cells or extremes. Retrieved from http://www.ncpublicschools.org/docs/data/management/policies/security/dmg-2009-004-se.pdf.

North Carolina Department of Education. (2020). School report cards. Retrieved from https://www.dpi.nc.gov/data-reports/school-report-cards

Ohno-Machado, L., Vinterbo, S., & Dreiseitl, S. (2002). Effects of data anonymization by cell suppression on descriptive statistics and predictive modeling performance. *Journal of the American Medical Informatics Association*, *9*(6 Suppl 1), s115–s119.

Oregon Department of Education. (2016). A summary to the Legislature of the annual report to the Legislature on English language learners 2014-2015 Oregon Department of Education. Retrieved from https://www.oregon.gov/ode/reports-and-data/LegReports/Documents/ell-report-summary-1415-final.pdf.

Raghunathan T. E., Reiter J. P., & Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. Journal of Official Statistics, 19(1), 1–16

Reiter, J. (2005a). Estimating Risks of Identification Disclosure in Microdata. *Journal of the American Statistical Association, 100*(472), 1103–1112.

Reiter, J. P. (2005b), Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Journal of the Royal Statistical Society – Series A, 168, 185–205.

Reiter, J. P. (2019). Differential privacy and federal data releases. *Annual Review of Statistics and Its Application*, 6(1), 85–101.

Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics & Data Analysis*, 53(4), 1475–1482.

Reiter, J. P. & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462–1471.

Schochet, P. Z. (2020). Analyzing Grouped Administrative Data for RCTs Using Design-Based Methods. *Journal of Educational and Behavioral Statistics*, *45*(1), 32–57.

Seastrom, M. (2010). Statistical methods for protecting personally identifiable information in aggregate reporting. SLDS Technical Brief. U.S. Institute for Education Sciences.

Singh, A. C., Yu, F., & Dunteman, G. H. (2004). MASSC: A new data mask for limiting statistical information loss and disclosure. In *Work Session on Statistical Data Confidentiality 2003*, Linden, H., Riecan, J., & Belsby, L. (Eds.). Eurostat, Luxemburg. Monographs in Official Statistics, 373–394.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2001). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society – Series A, (Statistics in Society)*, 174(2), 369–386.

Tipton, E. (2013) Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38, 239–266.

Tipton, E., & Olsen, R. B. (2018). A Review of Statistical Methods for Generalizing From Evaluations of Educational Interventions. *Educational Researcher*, 47(8), 516–524.

West, B.T. and Little, R. J. A. (2013). Nonresponse adjustment of survey estimates based on auxiliary variables subject to error. *Journal of the Royal Statistical Society – Series C (Applied Statistics)*, 62(2), 213–231.

Wisconsin Department of Public Instruction. (2018). Student Privacy. Retrieved from https://dpi.wi.gov/assessment/student-privacy.

State	# Students		# Schools		# Districts	
	Grade 4	Grade 8	Grade 4	Grade 8	Grade 4	Grade 8
ST 1 (2012)	79,382	76,976	1,142	779	424	404
ST 2 (2010)	50,946	49,177	774	442	---	---
ST 3 (2012)	71,263	74,081	1048	649	307	292
ST 4 (2010)	120,003	112,903	1,418	716	200	202
ST 5 (2010)	6,679	6,849	268	195	176	168
ST 6 (2014)	75,305	76,778	1,248	693	294	280
ST 7 (2009)	60,083	61,999	1,133	668	427	428
ST 8 (2012)	20,727	20,717	423	205	57	57

Table 1. *This table summarizes the datasets used in this investigation. These datasets are cross sections of eight statewide data systems for fourth and eighth grade.*

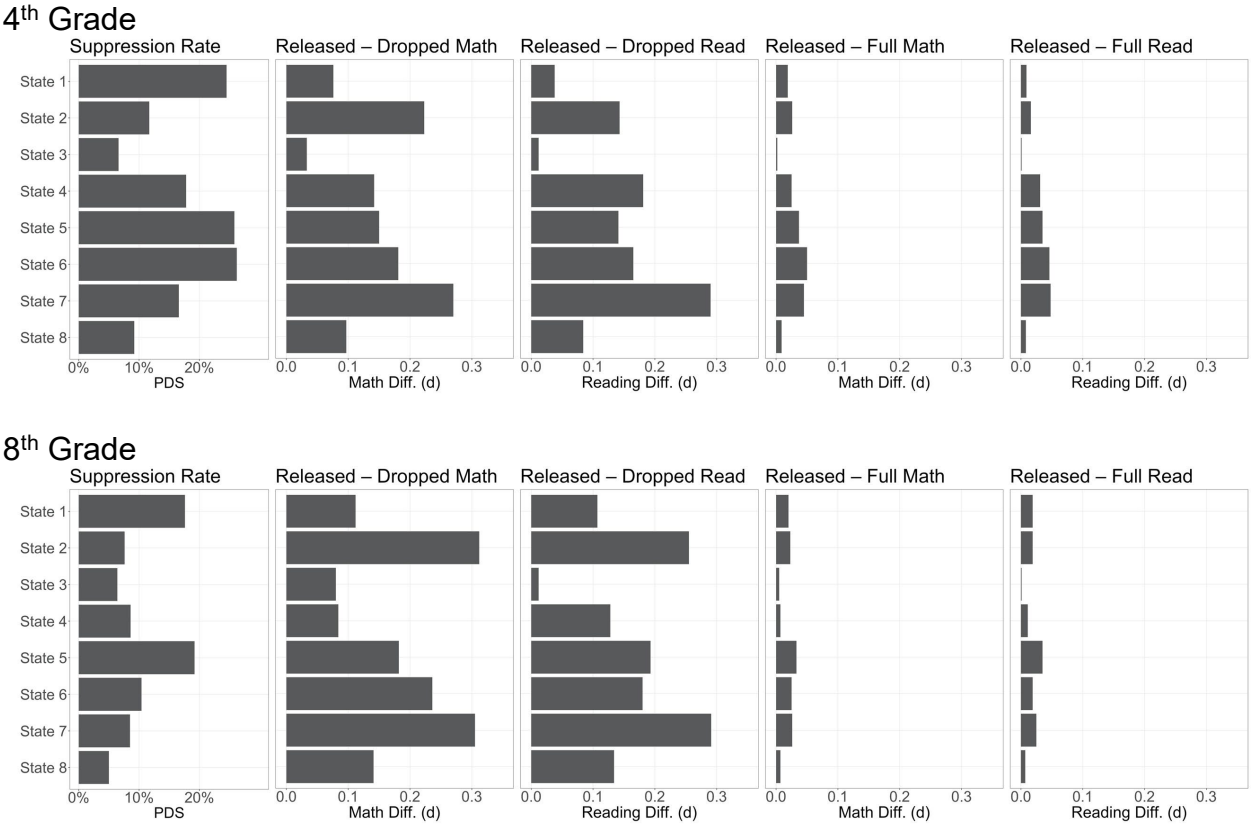

Figure 1. *These plots show the suppression rate (left most plot), the difference between released and dropped test scores for math and reading, and the difference between released and full test scores for math and reading. Each bar represents one state, plots on the top row correspond to 4th grade students, plots on the bottom row are for 8th grade students. Differences are reported in standard deviation units (SD).*

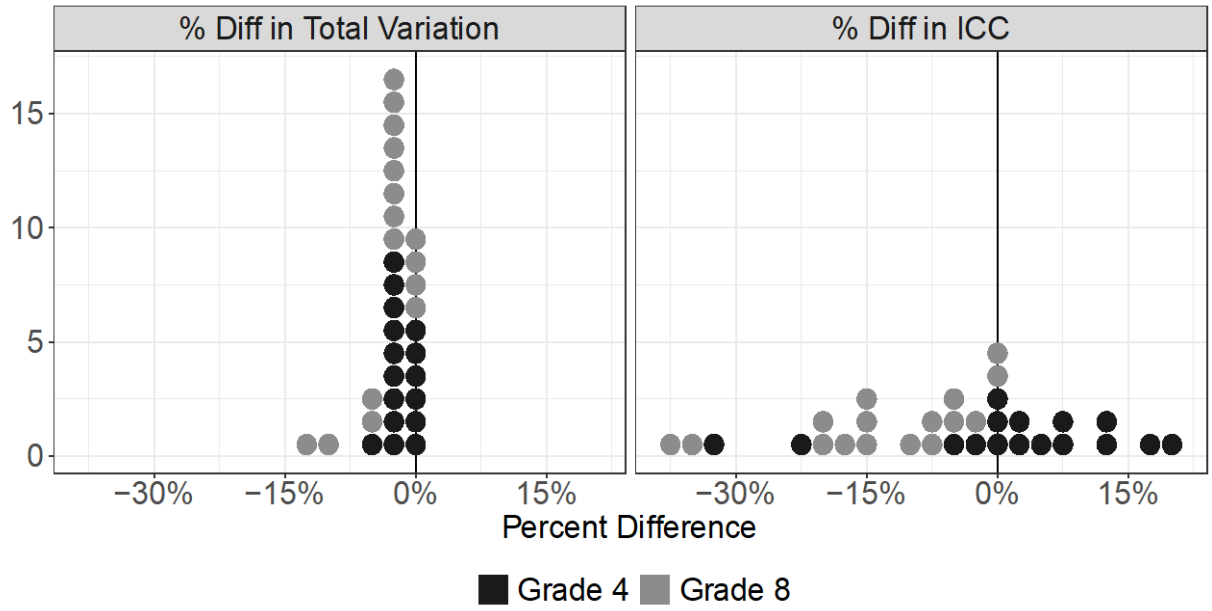

Figure 2. *These plots show the differences between the full and released data variance (left plot) and intra-class correlation (ICC) (right plot). Differences are reported at percent differences, and the dots correspond to a difference computed on a single cross section of data for a single test (math or reading). Dots are shaded by the grade of students in that cross section.*

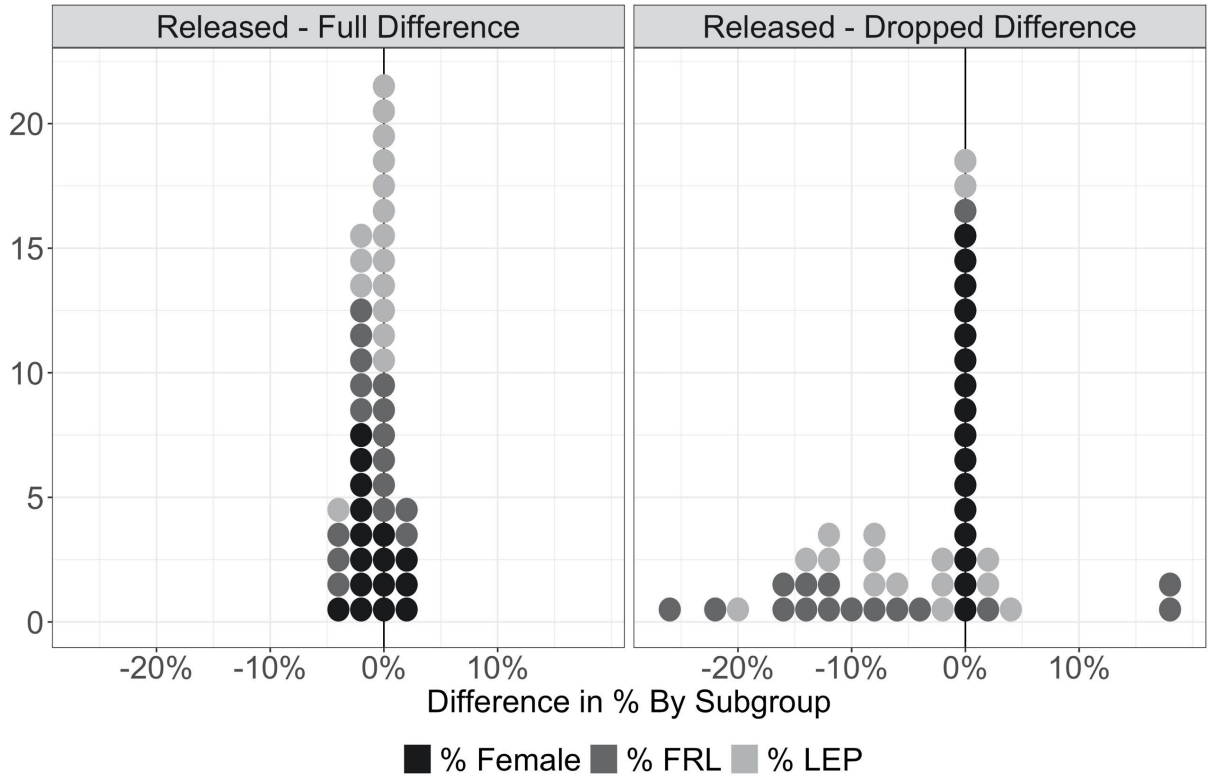

Figure 3. *These plots show the differences between the demographics in the released versus the full data (left plot) and the released versus dropped data (right plot). Dots correspond to the difference in the percentage of students in a demographic subgroup in the released data versus the full or dropped data computed on a single cross section of data. Differences are reported as a raw difference. Dots are shaded by subgroup: % female, % of students receiving free or reduced priced lunch (FRL), and % of students who are deemed to have limited English proficiency (LEP).*

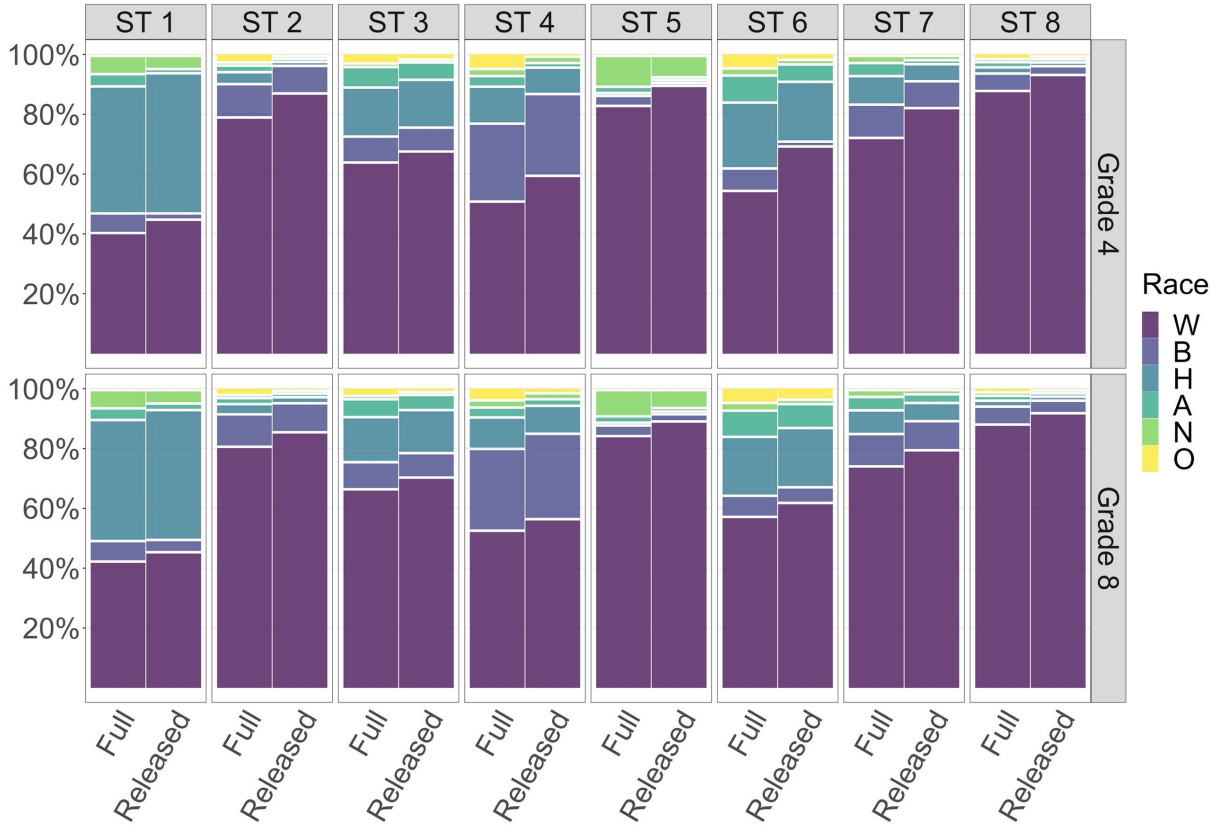

Figure 4. *This figure plots the differences in the racial composition of data released after microsuppression and the full data. Bars are shaded by racial subgroup: W = white, B = black, H = Hispanic, A = Asian, N = Native American, O = other.*

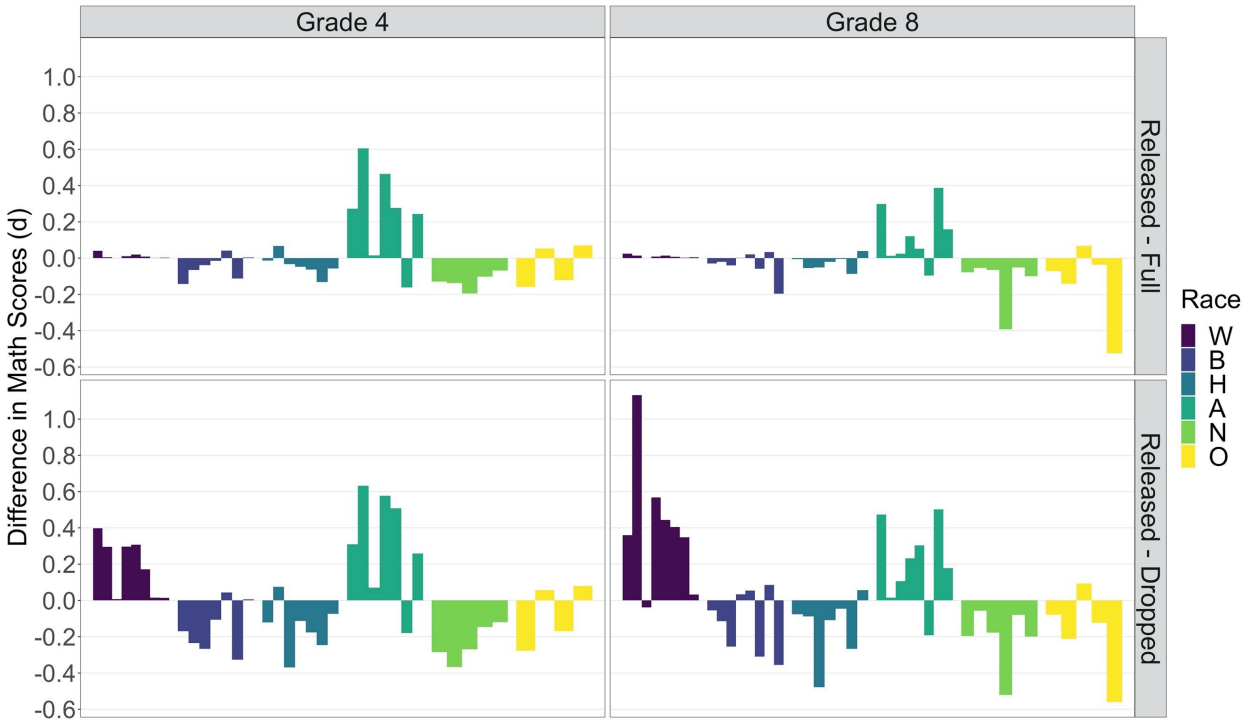

Figure 5. *This figure plots the difference in mean math test scores within racial subgroups for data released after microsuppression versus the full data and data dropped as a result of microsuppression. Each plot shows the difference in means on the scale of standard deviation units, and bars are shaded according to racial subgroup: W = white, B = black, H = Hispanic, A = Asian, N = Native American, O = other.*

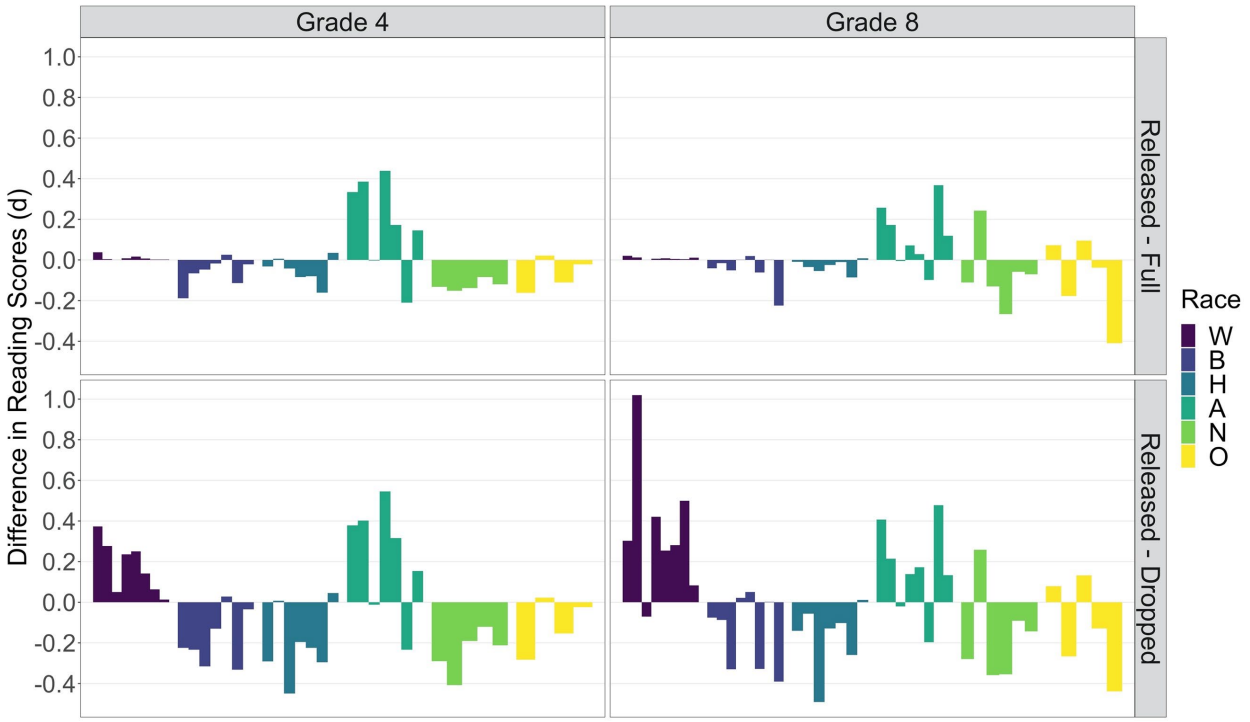

Figure 6. *This figure plots the difference in mean reading test scores within racial subgroups for data released after microsuppression versus the full data and data dropped as a result of microsuppression. Each plot shows the difference in means on the scale of standard deviation units, and bars are shaded according to racial subgroup: W = white, B = black, H = Hispanic, A = Asian, N = Native American, O = other.*

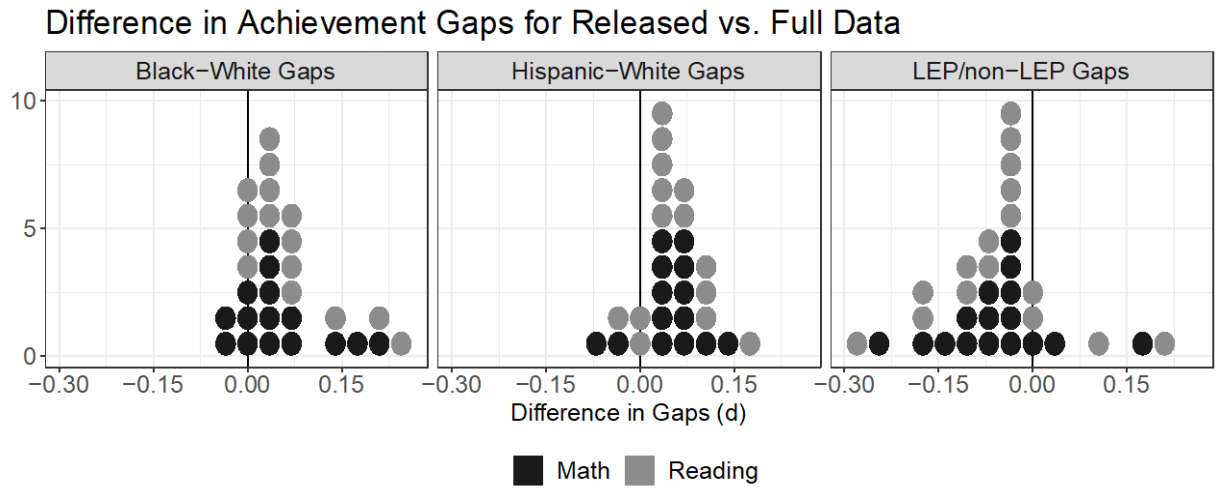

Figure 7. *These plots show the difference in achievement gaps in the data released after microsuppression versus the full data. Differences are reported on the scale of standard deviation units; positive values indicate the released data overstate the achievement gap. Dots correspond to differences computed on a single cross section, and they are shaded according to the state assessment subject on which they were computed: mathematics or reading. Vertical lines indicate a difference of zero; note that each plot as different x-axis scales.*

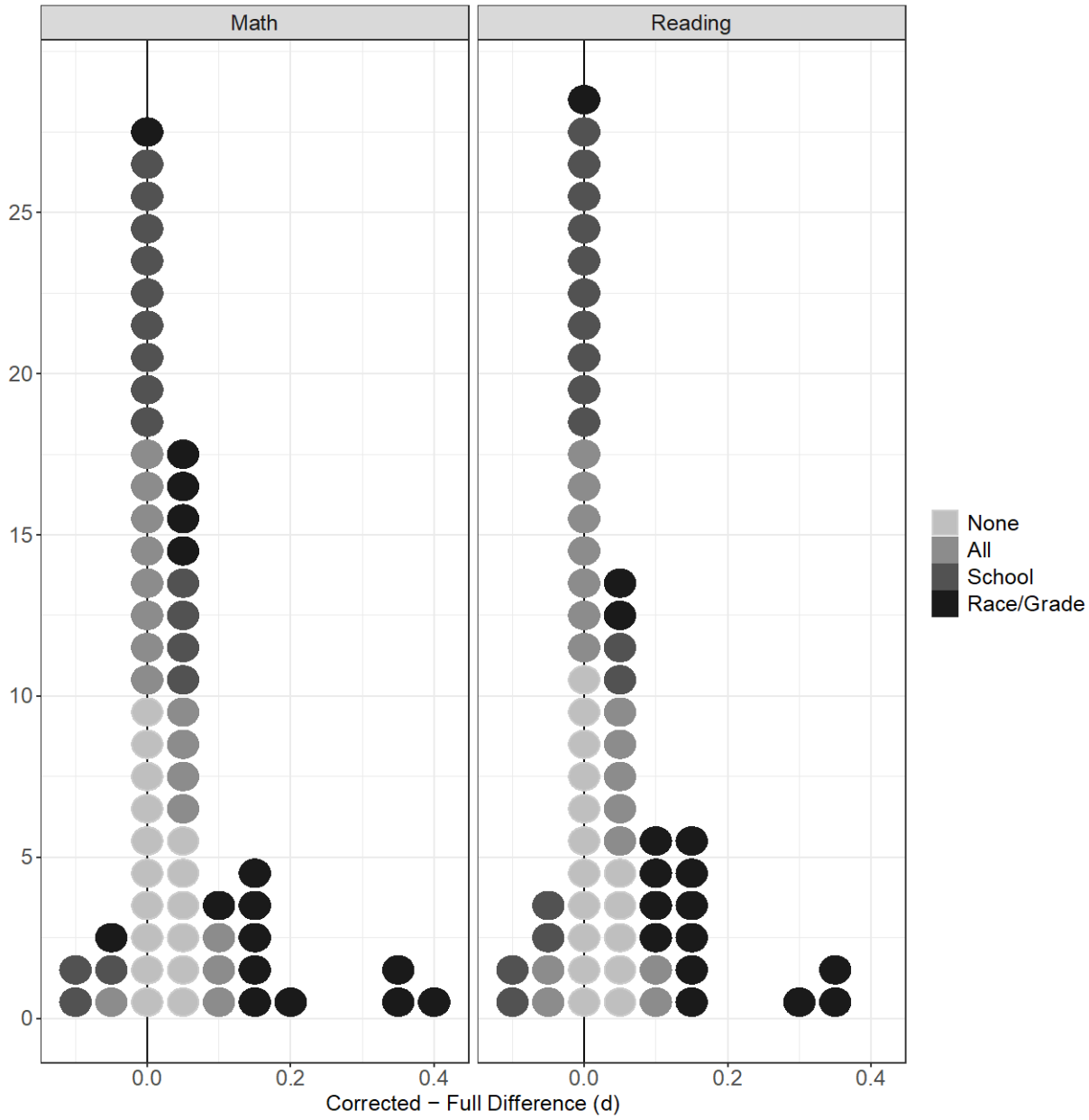

Figure 8. *This figure shows the standardized difference between the corrected average test score in the released data versus the actual average in the complete data. Differences are shown in standard deviation units. The plots are broken out by subject (math and reading), and dots correspond to a given cross section. Dots are shaded according to the variables used to compute post-stratification weights. "None" indicates no correction; "Race/Gender" weights by the joint proportions for race-gender combinations; "School" weights by school size; and "All" uses weights according to race, gender, and school.*