The effects of microsuppression on state education data quality

Jacob M. Schauer[1], Arend M. Kuyper[1], Eric C. Hedberg[3], & Larry V. Hedges[1]

1. Department of Statistics, Northwestern University
2. National Opinion Research Center (NORC)

Abstract

States often turn to a data masking procedure called microsuppression in order to reduce the risk of disclosing student records when sharing data with external researchers. This process removes records deemed to have high risk for disclosure should they be released. However, this process can lead to analyses that differ from those conducted on the complete (unmasked) data, especially if the records that are released reflect different types of students than those that are suppressed. This paper assesses the extent to which microsuppression can bias parameter estimates, and finds that while marginal test score means tend to be preserved in the masked data, conditional means for subgroups can exhibit bias as large as 0.3 standard deviations.

Balancing Utility & Security

Data collected by state education systems offer great promise for informing improvement in schools (Honig & Coburn, 2008; Conaway, et. al., 2015). Important insights often arise from analyses of these data conducted by external researchers, for instance from universities or research institutes. This presents something of a tradeoff for the state data systems, which are legally required to protect student privacy under laws such as the Family Educational Rights and Privacy Act (FERPA). States must balance the potential utility of releasing data against the risks of violating students' privacy. If data are never released, then states can largely avoid any disclosures of student data that could constitute a violation of FERPA. However, doing so will deny states of useful information about how their schools are operating. Conversely, if data are released too frequently, there is no guarantee of greater benefits, but considerably more risk of FERPA violations.

In an attempt to balance these two considerations, states have often 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 follows from 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, with some certainty, the rest of that student's record, including any sensitive information in that record (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 less 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 considerably 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. Tables published in journals or on websites of state departments of education are required to suppress values in cells with small counts or extreme values (e.g., North Carolina, 2009). Not surprisingly, this also occurs with public-use datasets available through state departments of education, which can contain student-level observations such as demographics, the school they attend, and their test scores. To protect against disclosures from public-use datasets, states suppress records for students in small schools or who may easily identified in larger schools (e.g., Massachusetts, 2014). However, we have engaged in projects where these rules have also been applied to secure releases of student-level data under data use agreements (DUA). In other words, state data administrations have applied masking and suppression procedures to many of the available 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 released. Analyses conducted on masked (microsuppressed) data may differ from those conducted on the complete (non-suppressed) data,

since the latter incorporates observations not used in the former. If they are substantially different, then inferences based on the released data may be inaccurate, and decisions based on them may not be optimal. 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 the validity of statistical inferences based on microsuppressed data. The following section provides a theoretical overview that describes potential inaccuracies in such inferences. We argue that it can be important not just how inaccurate analyses are, but also why. Then, using cross sectional data from states we provide empirical evidence about biases in estimates from microsuppressed data, as well as insight into what might be driving those biases. 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 more bias. Moreover, we demonstrate that in some instances, potential corrections, such as raking and post-stratification, may actually do more harm than good.

Microsuppression as Missing Data

Methods aimed at masking sensitive data are used by various statistical agencies, including state education data systems. A basic method was described by Duncan and Lambert (1986) as "cell suppression," and what others have termed "microsuppression," wherein records at high risk of disclosure are removed from the data prior to its release to external researchers. This is a seemingly common approach to data masking for state education data systems (Levesque, et. al., 2015), however the analyses of data masked in this way may not match those conducted on the full data. The extent to which they do can depend on the nature of the data and the way cells are determined. In this section, we provide a formalized description of the practice

and how states appear to be using it, as well as a theoretical discussion of how it can affect statistical inferences.

The risk of disclosing sensitive data can be understood within the framework of Duncan and Lambert (1986, 1989), who describe someone seeking sensitive information on a specific student (or set of students) as an "intruder." The intruder may know some information about that student, such as their grade, gender, etc., which Reiter (2005) denotes as a vector ***t***. The intruder attempts to match the information in ***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 ***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. Thus, the risk of disclosing sensitive information is high if an intruder can match their information to only a few records.

To protect against this, microsuppression works by deleting records with a high risk of disclosure; that is, records for which a potential intruder's vector of information ***t*** has only a few matches. How this occurs, and how it can affect statistical inferences 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. The general guidance for state data systems is to protect personally identifiable information, such as names or addresses, but also information that can be combined to identify individuals (Johnson, 2007). A common step for state data systems involves removing easily identifying features such as names and social security numbers. However, as described above, an intruder could potentially use some of the remaining fields in a table (e.g., a student's school, grade, race, and gender) in concert to potentially identify a student.

It appears that states have adopted the idea that observations in small cells ought to be deleted (Johnson, 2007; Levesque, et. al., 2015). Restrictions are in place for published summaries of state data, such as in tables that present summary statistics like counts or mean test scores of students in a state for a given set of categorical demographic variables (e.g., race, gender, or if the student scores proficient on a state exam). In these cases, a common approach appears to be that if cells contain fewer than six individuals, then no statistic is reported (e.g., Oregon, 2016; Wisconsin, 2018). However, these rules are also applied to raw student-level data in addition to summary tables. If the data contain test scores and categorical demographic variables, then cells are often determined by those demographic variables, and records in small cells (often fewer than 10 students) are deleted (e.g., Massachusetts, 2014; Montana, 2018; Nebraska, 2013; North Carolina, 2009). While this is applied to public use datasets available from states, we have experienced data exchanged under secure DUAs subject to such masking.

Since observations get deleted in the process of microsuppression, analyses conducted on the masked data may not match the results of the same analyses conducted on the complete (unmasked) data. This issue has been noted by researchers in different fields as a shortcoming of microsuppression (Kelly, 1992; Ohno-Machado, 2002; Matthews, et. al., 2017). Indeed, if there are large differences between analyses conducted on the complete and masked data, inferences obtained by researchers working with masked data will be inaccurate and lead to potentially poor policy moves. Thus, it is important to know if and when inferences from masked data may be considered accurate.

In the context of data masking, we might consider accurate analyses as those that return the same results when run on the complete and masked datasets. How we conceive of accuracy will depend on if the data comprise the entire population or are considered a sample. If the data

are considered an entire population, then population parameters in both the masked and complete datasets ought to be (nearly) identical. If instead the data are only a sample, then the properties of the analyses ought to be preserved. For instance, we might consider statistical estimates based on suppressed datasets valid if they are unbiased and have properly estimated standard errors.

Bias

Whether we consider the complete data as an entire population or as a sample, we can conceive of bias in analyses of masked data in terms of the population parameter μ_{ret} in the masked dataset and the (potentially different) population parameter μ in the complete dataset. If the complete data is the full population, then μ and μ_{ret} are merely the parameters computed in each dataset. If the data are considered a sample, then μ and μ_{ret} are the estimands of interest for the complete and masked dataset, respectively. A similar notation has been used in the literature on missing data to describe the differences between inferences with the complete data, and those involving datasets with missing values (see Little & Rubin, 2002).

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

$$\mu = E[Y|X = 1] \quad (1)$$

If we had the full dataset, we might compute (or if we have a sample, estimate) this with the conditional mean of test scores for minorities. This would give us the value of μ, or if the data are just a sample, this would constitute an unbiased estimate of μ. However, if suppression has occurred, we do not have access to the full dataset. Let S be an indicator for whether or not

an individual is in a small stratum. Then the masked (and ultimately released) data would comprise only observations for which $S = 0$. Thus, if we were to take the average test score for minorities in the masked data, it would be an unbiased estimate for

$$\mu_{ret} = E[Y|X = 1, S = 0] \quad (2)$$

The average test score for minorities in the masked data will not provide the value of μ (or is a biased estimator of μ) if $\mu_{ret} \neq \mu$. In other words, the mean in the full 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.

Whether we treat the full data as a complete population or as a sample, the difference between μ and μ_{ret} can be referred to as bias. This bias can be expressed as

$$Bias = (E[Y|X = 1, S = 0] - E[Y|X = 1, S = 1])P[S = 1|X = 1] \quad (3)$$

This expression contains two components. The first is the difference in expectation between the retained ($S = 0$) and suppressed ($S = 1$) test scores. 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 retained and suppressed observations. This can result in bias, but it also means that correcting for this bias may be difficult. This is because the masked data may not contain useful information about the observations that are dropped from the dataset prior to its release.

The second component is the probability that an observation is suppressed, what might be called a rate of suppression or the fraction of information missing (FIM). The FIM is an important quantity here. If it is 0%, then the bias is zero since no observations are suppressed. If it is larger, then there may be substantial bias. Moreover, if it is 100%, then we will be unable to estimate certain parameters. For example, suppose all of the minority students are suppressed in

a given state, since there are only a few in each school. Then we would be unable to estimate any parameter that involves minority students. Thus, the suppression rate can determine not only the extent of the bias, but also whether a parameter can be estimated at all.

Variance

Another type of parameter that may be of interest to researchers and policymakers is the variation in the data, and how this variation is partitioned at different organizational levels. Variance may be a parameter of interest in and of itself, however, it is also worth considering because the precision of many estimates used in education research are a function of it. Specifically, the precision of estimates will depend on the total variation and the intra-class correlation (ICC). The total variation refers to the population variance of a variable, such as the variance of all the math scores in the complete dataset. Greater total variation means that estimators tend to be less precise. The ICC quantifies the extent to which student test scores in the same school or district are correlated. If students in the same school have similar test scores, the ICC is high. The greater the ICC, the less precise estimators tend to be.

In this study, we consider school-level ICCs. Denote the variation of 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

$$\rho = \frac{\tau^2}{\tau^2+\sigma^2} \tag{4}$$

However, in the masked data, the relevant parameters are τ^2_{ret} and σ^2_{ret}, and the ICC in the masked data can be written as

$$\rho_{ret} = \frac{\tau^2_{ret}}{\tau^2_{ret}+\sigma^2_{ret}} \tag{5}$$

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 actual state education data. Our approach is to mimic the behavior of data masking procedures using complete cross sections of this data for each state. Using these cross sections, we can divide records into risk strata and suppress the small strata according to rules typically used by states when releasing student-level data. Analyses can then be run on the masked data as well as the complete data, and their results can be compared. This section details the data used, how it gets masked, and the analyses conducted.

We obtained data from eight states. For each of these states, we took completes 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. Since the data contain test scores, we delete records in strata smaller than 10. This follows the guidance of the privacy and data management plans of states described in the previous section. Moreover, by only stratifying on these three common variables, this experiment serves as an upper bound for the FIM, as increasing the number of strata for a given dataset can lead to a greater number of records being in small strata.

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 masked data to those of the complete data. To get a better sense of why differences emerge between masked and complete datasets, we examine suppression rates, as well as systematic differences between the suppressed (i.e., observations that are dropped) and retained data.

The analyses conducted treat the cross sections as complete populations, and so quantities computed are regarded as the value of population parameters. Often methods used in analyses of education data, such as maximum likelihood or least squares, obtain unbiased estimates of parameters. However, if such methods are applied to masked data, then the resulting estimates will be unbiased for the wrong parameters; they will be unbiased for μ_{ret} instead of μ. Thus, by treating the complete data as fixed populations, we can get a sense of the difference between these parameters for cross sectional data in education.

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:

$$\delta = \frac{\mu_{ret} - \mu_{full}}{\sigma_{full}} \quad (6)$$

where μ_{ret} and μ_{full} are the mean scores for the masked and complete data, respectively, and σ_{full} 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.

Table 2 shows the differences in test score means for each cross section of data. The "Masked – Full" columns display the standardized mean differences for math and reading scores. In these columns 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 FIM and the differences between the records that are retained versus those that are dropped. Table 2 provides insight into how these relate to state education data. What we see in the FIM column is that the proportion of observations that get suppressed varies greatly across states. Some are required to delete as much as a quarter of their data under the suppression rules, while others suppress only 5% of it. Moreover, the differences between the retained and suppressed records ("Masked – Dropped"

columns) 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 is larger than what might be considered a "small" 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 FIM and the Masked-Dropped differences are related to the Masked-Full differences, they are not related to each other in the data. Viewing the FIM and "Masked – Dropped" columns 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

The reported precision of analyses involving student test scores will frequently depend on the total variation of test scores, as well as the correlation between scores among students who are educated together, for instance in the same school. A measure of this correlation is the school-level ICC. If the total variation or the ICC are smaller in the masked data than in the full data, then analyses that treat the data as a sample will underestimate the uncertainty of estimates.

Table 3 compares the differences of these two quantities for test scores in the cross sections of data used in this article. For each grade, state, and test, it shows the percent difference in total variation ("Total Diff"); negative values indicate that the masked data exhibit less total variation than the complete data. What can be seen from these columns is that the masked 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 among State 2 eighth graders the total variation decreased by as much as 11.5%.

Table 3 also shows results for ICCs. For each state, test, and grade, the ICC is reported for the complete and masked data, and the percent difference between them is given in the "Diff" columns. 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 masked data have larger ICCs, for others it is negative, meaning that the ICCs are smaller in the masked data. This contrasts with the results for eighth graders, where the masked data almost universally have smaller ICCs, and in some instances (e.g., State 6 or State 8) substantially so.

Demographics

Demographic differences between the complete and masked data tend to follow two main patterns. Retained data largely contains fewer FRL and LEP students, as well as a smaller percentage of minorities when compared to the complete data. For some demographic variables, these differences can be modest, such as gender and LEP and FRL designations. Table 4 shows the differences between datasets for the proportion of LEP, FRL, or female students for a given dataset. It is organized by demographic variable, and shows the percentage of students in that demographic in the full, masked (retained), and suppressed (dropped) records. The differences for these variables tends to be less than 2% between the full and masked data. However, the students whose records get suppressed tend to be much more likely to receive FRL or be designated LEP.

More sizable differences occur for race. Figure 1 shows the proportion of each race present in each cross section for the complete data ("Full") and the masked data ("Masked") for

each state and grade. The figure shows that masked data tends to contain more white students and fewer black and Hispanic students. Differences in proportions can be as large as 16% (State 6, whites, fourth grade). Moreover, under these suppression rules, racial categories can disappear entirely from the masked data, as is the case in State 5.

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

Conditional Means

To assess how relationships between variables can be affected by microsuppression, we can examine conditional distributions of test scores for subgroups. For instance, we might determine if the mean test scores for students receiving FRL are the same in the masked and complete datasets. In that case, we would be interested in the difference between $E[Y \mid FRL, S = 0] - E[Y \mid FRL]$. If these types of differences are large, they will affect conditional inferences, such as in linear regression.

Figures 2 and 3 show how mean test scores for each race differ across datasets. The top panel of each figure plots the standardized mean difference between the masked versus complete data computed within each racial subgroup for each state, grade, and test. Each bar corresponds to a given state and grade, and bars above zero indicate that the masked conditional mean for that state and grade exceeds the conditional mean in the complete data, and bars below zero indicate that the masked data has a smaller mean than the complete data. What can be seen in both figures is that the masked data will tend overstate the achievement of white and Asian students, but

understate the achievement of black, Hispanic, and Native American students. These differences can be large in magnitude. For mean test scores among black and Hispanic students the difference is on the order of about 0.1 to 0.2 standard deviation units, but for Native American students, this difference is even larger.

These differences are driven both by the FIM for these subgroups, as well as differences in the retained versus dropped records. Recall from the previous section that a greater fraction of black, Hispanic, and Native American students are dropped during microsuppression, which means the FIM for these subgroups will be higher. Moreover, the lower panels of Figures 2 and 3 show the difference in average achievement between individuals whose records are retained 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 retained score on average about 0.2 standard deviations lower than those who are suppressed. Meanwhile, the white and Asian student who are retained tend to score 0.2-0.5 standard deviations higher than those who are suppressed.

Taken together, the masked data tend to contain higher achieving white students and lower achieving black and Hispanic students. This will lead to greater racial achievement gaps in the masked data relative to the complete data. Tables 5 and 6 show the differences for each state and grade for the black-white (Table 5) and Hispanic-white (Table 6) gaps. These tables show the standardized difference in mean test scores between white and minority students in the complete and masked data, and the difference between those gaps in the "Diff" columns. Positive values in the "Diff" column correspond to masked 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 both racial achievement gaps.

Also present in those tables are blank cells, which are there to indicate that all of the Hispanic student records were suppressed. This occurs with other racial groups, including Native Americans. Suppression conducted according to these rules, then, can render certain analyses involving these racial subgroups impossible in masked data.

For the other demographic variables in the data, differences in achievement gaps between complete and masked data can be 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 masked data, but only by about 0.01 standard deviations in magnitude. However, the achievement gap between non-LEP and LEP students may be substantially different in the complete and masked data. Table 7 shows the differences between mean test scores for non-LEP and LEP students in the complete and masked data, as well as the difference between these gaps ("Diff" columns); values are computed on the scale of standardized mean differences as in equation (6). Note that for math scores, the masked data will often overstate the achievement gap (i.e., the difference between the complete and masked 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 when the fraction of LEP students who get suppressed ("FIM" column) is larger.

Raking & Post-Stratification

One potential correction that can be used with missing observations in education data is to re-weight observations when computing statistics. For instance, if after suppression, there are

fewer Hispanic female students in the data, then one might be tempted to upweight observations that correspond to Hispanic females. Precisely how these observations should be weighted requires some knowledge about how many Hispanic females should be in the data. This type of information is often available in administrative reports, for instance from the National Center for Education Science's Common Core of Data (CCD).

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 z as $P[\boldsymbol{Z} = z]$, and reweight observations in the masked data by $1/P[\boldsymbol{Z} = z]$. 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 Table 8. For each test, the table shows the bias of the unweighted mean (the "None" column), which is how Table 2 is computed, and a mean that weights observations by race and gender; by school; and by race, gender, and school (the "All" column). Values are reported in terms of standardized mean differences as computed by equation (6); positive values indicate that the masked-data mean is larger than the complete-data mean.

What can be seen in the table is that incorporating race and gender into the weights can greatly exacerbate bias. The differences in those columns are considerably larger than those in the unweighted column, meaning that less bias is obtained by not weighting by race or gender. For some states and grade, 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. Finally, weighting by all three increases the bias relative to weighting just by school size.

This increase in bias can be explained by the results in Figures 2 and 3 from the previous section. Recall that these showed that the difference between retained and dropped observations exhibit substantial differences within racial subgroups. This is particularly true for black and Hispanic students. At the same time, from Figure 1, we see that a greater proportion of black and Hispanic students are removed from the data under the suppression rules. Thus, the black and Hispanic students who remain in the masked data will receive greater weight, however, they will also differ substantially from the records of black and Hispanic students that are dropped.

Conclusions

This paper has sought to quantify the effect of masking state education data via microsuppression. Using complete cross sections of state data, we have attempted to do so by applying seemingly common suppression rules to these data and comparing quantities computed in the complete and masked data. While the data and suppression rules used in this study do not cover all possible scenarios, they do offer some idea of what might be possible of masked data.

What we find is that marginal means of test scores appear minimally affected by masking. Moreover, while the total variation in test scores tends to be smaller in masked data, it is not usually vastly smaller. As well, school-level ICCs tend to decrease slightly. In sum, the marginal distributions of test scores in masked data, while not identical to those of the complete data, appear to preserve the values of some population parameters, and do so while limiting the risk that individual students are identified.

However, conditional distributions tend to exhibit large differences between the complete and masked data. The achievement of white and Asian students, for instance, tends to be higher in the masked 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.

There are two main reasons for these differences. The first is the FIM, which can be large for some subgroups. The second is that the data that gets suppressed tends to contain higher achieving black and Hispanic students and lower achieving white and Asian students. As well, the suppressed observations are more likely to include students who have LEP or receive FRL, and those students achievement tend to be higher than similar students whose data are retained.

Because there are such large difference between suppressed and retained data, particularly within racial subgroups, methods that attempt to correct marginal mean estimates by post-stratifying by race will exacerbate bias. Whereas unweighted marginal mean test scores in the masked data are typically within 0.05 standard deviations of the complete-data mean, when observations are weighted according to race or gender the difference between the weighted mean and the complete-data mean can be larger than 0.2 standard deviations.

In addition, for most analyses conducted in this study, the extent to which quantities are preserved in masked data can vary substantially by state and grade. Racial achievement gaps, for instance, may be quite similar between the masked and complete data for one state, and very different for another. Moreover, to the extent that general patterns exist in differences between the complete and masked data (i.e., most race gaps are overstated in the masked data), the magnitude of those differences can vary by state.

Because microsuppression often does not reliably preserve important quantities and relationships in state education data, we would urge caution in interpreting analyses based on

data masked in this way. This calls into question just how much utility there is in microsuppressed data, and how much of a balance is actually struck by state data systems. A potential consideration for states and researchers alike is whether more sophisticated masking methods, such as synthetic data methods used by other government agencies, might better preserve relationships in education data.

References

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.

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.

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.

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.

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

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

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., & 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.

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.

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

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

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

Table 1: State Data Cross Sections.

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 2: Differences in Mean Test Scores. This table shows the effects of microsuppression on state average test scores for math and reading. For each state and grade, the proportion of data that gets suppressed (FIM), differences in test scores between the masked and complete data (Masked – Full), and differences between the masked data and the data that gets dropped as part of microsuppression (Masked – Dropped) are reported on the scale of standardized mean differences.

			Masked – Full		Masked - Dropped	
State	Grade	FIM	Math	Reading	Math	Reading
ST 1	4	24.5%	0.019	0.009	0.076	0.038
	8	17.6%	0.020	0.019	0.112	0.107
ST 2	4	11.7%	0.026	0.016	0.223	0.143
	8	7.6%	0.023	0.019	0.312	0.255
ST 3	4	6.6%	0.002	0.001	0.033	0.012
	8	6.4%	0.005	0.001	0.080	0.012
ST 4	4	17.8%	0.025	0.031	0.142	0.181
	8	8.6%	0.007	0.011	0.084	0.128
ST 5	4	25.8%	0.037	0.035	0.150	0.141
	8	19.2%	0.033	0.035	0.182	0.193
ST 6	4	26.2%	0.050	0.046	0.181	0.165
	8	10.4%	0.025	0.019	0.236	0.180
ST 7	4	16.6%	0.045	0.048	0.270	0.290
	8	8.5%	0.026	0.025	0.305	0.291
ST 8	4	9.2%	0.009	0.008	0.097	0.084
	8	5.0%	0.007	0.007	0.141	0.134

Table 3: Variance Components for Test Scores. This table shows differences in test score variation between the complete and masked data. It includes columns for the total variation, and school-level ICCs for both math and reading achievement across fourth and eighth grade.

		Grade 4				Grade 8			
State	Test	Total Diff	Full ICC	Masked ICC	Diff	Total Diff	Full ICC	Masked ICC	Diff
ST 1	Math	-2.2%	0.177	0.176	-1.0%	-1.7%	0.202	0.195	-3.1%
ST 1	Reading	-0.7%	0.182	0.194	6.5%	-2.4%	0.177	0.168	-5.2%
ST 2	Math	-1.7%	0.152	0.151	-0.7%	-11.5%	0.250	0.155	-38.1%
ST 2	Reading	-1.4%	0.133	0.133	-0.2%	-10.4%	0.235	0.151	-35.6%
ST 3	Math	-0.5%	0.118	0.122	3.4%	-0.9%	0.142	0.143	0.4%
ST 3	Reading	-0.2%	0.132	0.136	3.0%	-0.6%	0.129	0.129	0.2%
ST 4	Math	0.3%	0.155	0.182	17.5%	-5.9%	0.241	0.197	-18.4%
ST 4	Reading	0.8%	0.140	0.170	21.2%	-2.7%	0.184	0.166	-10.1%
ST 5	Math	-4.2%	0.146	0.112	-23.4%	-0.9%	0.107	0.099	-8.0%
ST 5	Reading	-1.8%	0.098	0.068	-31.4%	0.3%	0.063	0.054	-15.3%
ST 6	Math	-1.6%	0.211	0.222	4.8%	-4.0%	0.202	0.173	-14.3%
ST 6	Reading	-1.6%	0.154	0.167	8.2%	-3.3%	0.133	0.107	-19.5%
ST 7	Math	-0.5%	0.170	0.189	11.7%	-2.9%	0.237	0.228	-3.9%
ST 7	Reading	-2.6%	0.138	0.155	12.2%	-3.4%	0.183	0.171	-6.8%
ST 8	Math	-2.2%	0.036	0.035	-4.2%	-2.3%	0.027	0.023	-13.8%
ST 8	Reading	-2.2%	0.044	0.042	-3.7%	-2.1%	0.035	0.028	-20.6%

Table 4: Demographics: FRL, LEP, and Gender

		4th Grade			8th Grade		
Variable	State	Full	Masked	Dropped	Full	Masked	Dropped
LEP	ST 1	9.2%	9.9%	7.0%	1.7%	1.3%	3.8%
	ST 2	2.5%	0.8%	15.5%	1.4%	0.7%	8.7%
	ST 3	9.1%	9.1%	7.8%	5.0%	5.0%	4.6%
	ST 4	7.8%	5.5%	18.1%	5.8%	5.0%	13.9%
	ST 5	0.9%	0.4%	2.4%	0.9%	0.8%	1.3%
	ST 6	20.6%	19.9%	22.6%	17.3%	17.6%	14.5%
	ST 7	7.1%	3.9%	23.0%	5.1%	4.1%	16.4%
	ST 8	0.8%	0.3%	5.7%	0.6%	0.2%	7.8%
FRL	ST 1	57.8%	58.0%	56.9%	53.7%	53.6%	54.3%
	ST 2	50.6%	52.7%	34.2%	54.5%	55.8%	38.5%
	ST 3	37.5%	37.3%	40.4%	35.8%	35.3%	43.5%
	ST 4	53.5%	50.9%	65.1%	49.2%	48.2%	60.1%
	ST 5	29.3%	25.1%	41.5%	24.2%	21.1%	36.9%
	ST 6	49.3%	46.4%	57.3%	45.0%	44.4%	50.8%
	ST 7	37.6%	33.3%	58.9%	33.2%	31.3%	53.8%
	ST 8	56.8%	55.7%	67.5%	52.3%	51.6%	66.1%
Female	ST 1	49.9%	49.1%	48.8%	50.0%	49.0%	48.8%
	ST 2	50.9%	48.4%	48.1%	47.0%	48.1%	48.2%
	ST 3	51.3%	48.8%	48.6%	50.4%	48.7%	48.6%
	ST 4	50.1%	49.0%	48.7%	49.9%	48.8%	48.7%
	ST 5	47.4%	48.7%	49.1%	52.7%	48.7%	47.7%
	ST 6	49.8%	48.7%	48.2%	49.3%	48.7%	48.6%
	ST 7	50.1%	48.6%	48.3%	48.2%	48.7%	48.8%
	ST 8	48.9%	48.5%	48.4%	46.3%	49.0%	49.2%

Figure 1: Racial Composition of Datasets

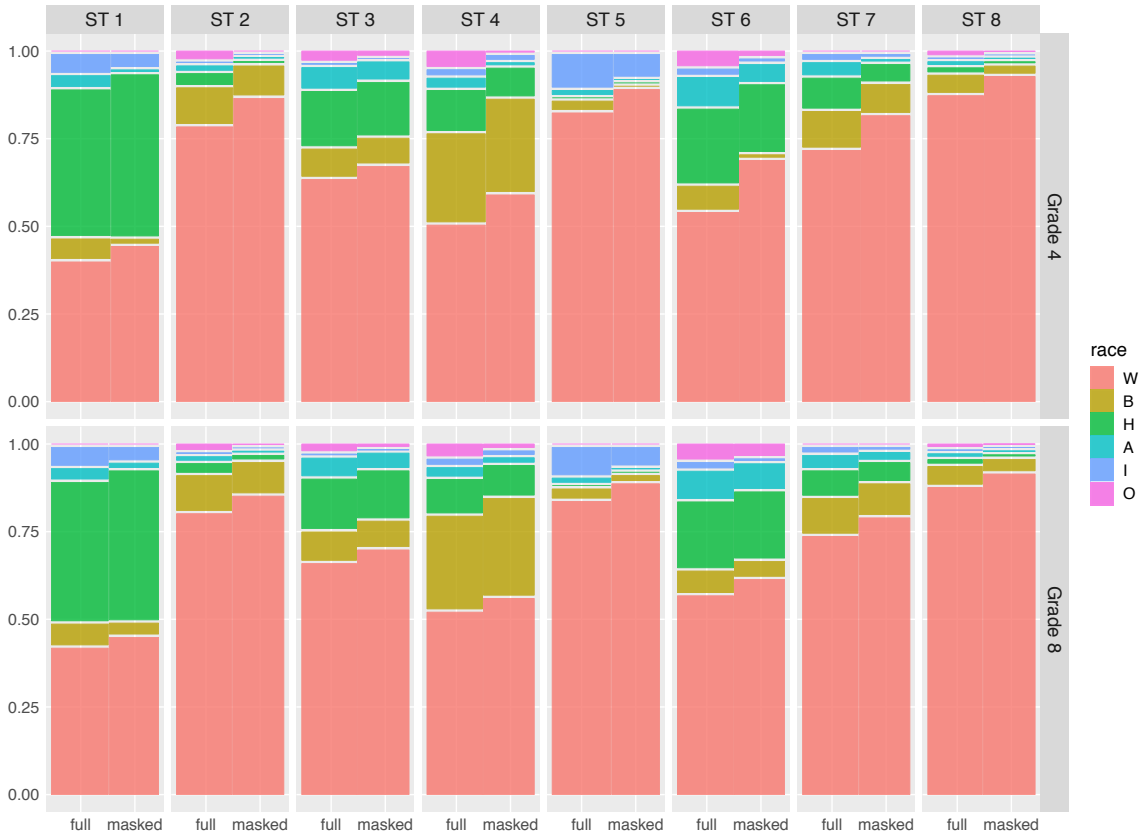

Figure 2: Racial Subgroup Differences in Math

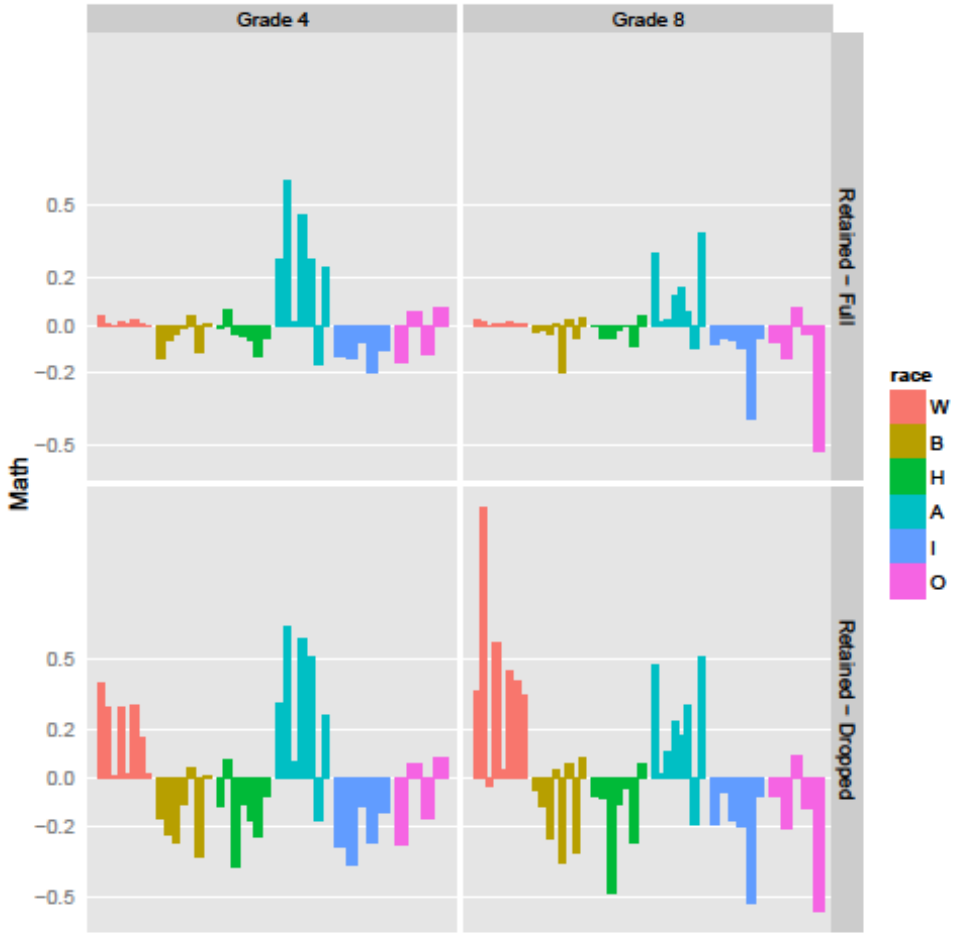

Figure 3: Racial Subgroup Differences in Reading

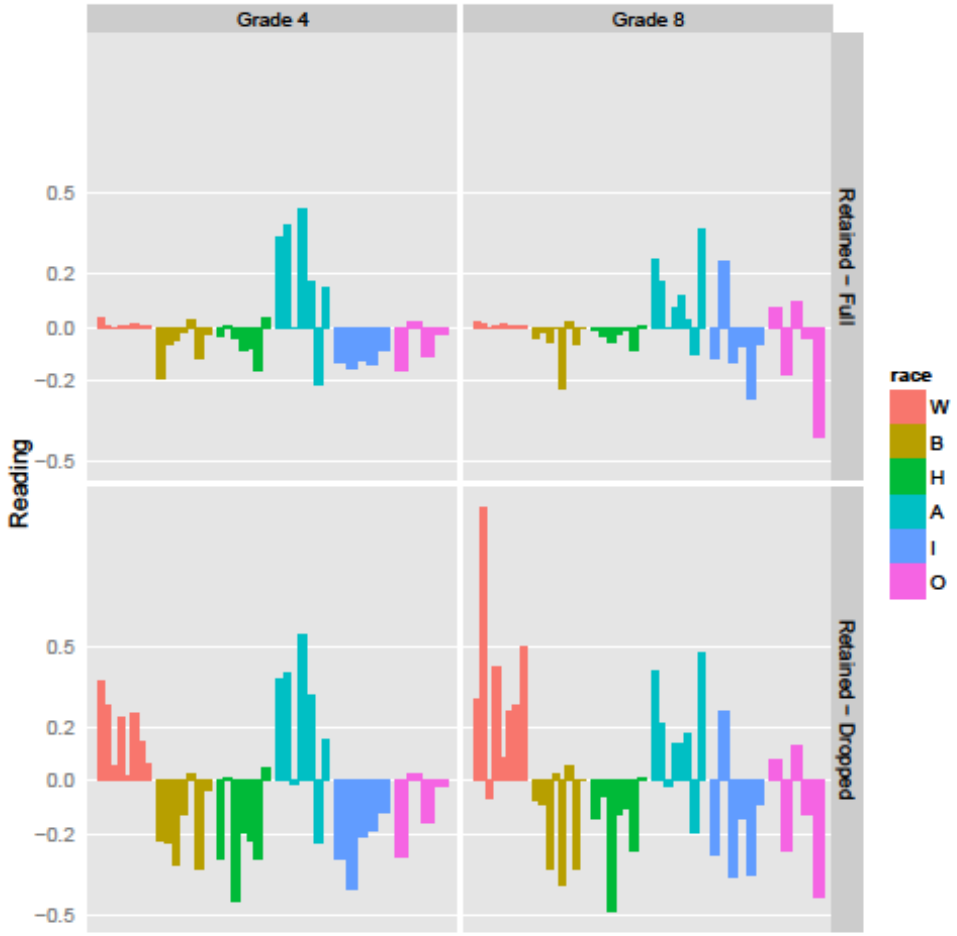

Table 5: Black-White Achievement Gap Differences. For each state and grade, this table shows the black-white achievement gap in math and reading achievement as computed in the full dataset, the masked data, and the difference between these gaps (Diff) in the masked versus full data.

		Math			Reading		
State	Grade	Full	Masked	Diff	Full	Masked	Diff
ST 1	4	0.582	0.755	0.174	0.614	0.830	0.216
ST 1	8	0.574	0.625	0.051	0.534	0.595	0.061
ST 2	4	0.692	0.763	0.072	0.557	0.632	0.075
ST 2	8	0.692	0.725	0.033	0.531	0.559	0.028
ST 3	4	0.906	0.949	0.043	0.768	0.812	0.044
ST 3	8	0.805	0.844	0.039	0.693	0.747	0.054
ST 4	4	0.797	0.820	0.023	0.786	0.810	0.024
ST 4	8	0.752	0.758	0.007	0.784	0.789	0.005
ST 5	4	0.613	--	--	0.502	--	--
ST 5	8	1.021	1.239	0.218	0.710	0.951	0.240
ST 6	4	0.098	0.075	-0.023	0.092	0.082	-0.010
ST 6	8	0.070	0.063	-0.007	0.079	0.068	-0.011
ST 7	4	0.936	1.063	0.126	0.889	1.017	0.129
ST 7	8	1.058	1.126	0.067	0.943	1.013	0.070
ST 8	4	0.243	0.239	-0.003	0.169	0.193	0.025
ST 8	8	0.204	0.168	-0.036	0.183	0.186	0.003

Table 6: Hispanic-White Achievement Gap Differences. For each state and grade, this table shows the Hispanic-white achievement gap in math and reading achievement as computed in the full dataset, the masked data, and the difference between these gaps (Diff) in the masked versus full data.

		Math			Reading		
State	Grade	Full	Masked	Diff	Full	Masked	Diff
ST 1	4	0.506	0.559	0.053	0.637	0.703	0.066
ST 1	8	0.518	0.547	0.029	0.561	0.590	0.028
ST 2	4	0.359	0.295	-0.064	0.213	0.211	-0.002
ST 2	8	0.299	0.365	0.066	0.217	0.265	0.047
ST 3	4	0.774	0.810	0.036	0.766	0.804	0.038
ST 3	8	0.857	0.908	0.052	0.822	0.881	0.059
ST 4	4	0.602	0.658	0.056	0.802	0.889	0.088
ST 4	8	0.550	0.578	0.028	0.761	0.792	0.031
ST 5	4	--	--	--	--	--	--
ST 5	8	--	--	--	--	--	--
ST 6	4	0.536	0.612	0.077	0.568	0.658	0.090
ST 6	8	0.520	0.538	0.019	0.490	0.509	0.018
ST 7	4	0.632	0.769	0.137	0.711	0.887	0.176
ST 7	8	0.756	0.853	0.097	0.707	0.803	0.096
ST 8	4	0.211	0.269	0.058	0.179	0.145	-0.034
ST 8	8	0.055	0.015	-0.040	0.056	0.052	-0.004

Table 7: LEP Achievement Gap Differences. For each state and grade, this table shows the achievement gap between LEP and non-LEP students in math and reading achievement as computed in the full dataset, the masked data, and the difference between these gaps (Diff) in the masked versus full data.

			Math			Reading		
State	Grade	FIM	Full	Masked	Diff	Full	Masked	Diff
ST 1	4	18.5%	-0.981	-1.033	-0.052	-1.175	-1.198	-0.023
ST 1	8	38.8%	-1.175	-1.324	-0.150	-1.662	-1.732	-0.071
ST 2	4	72.9%	-0.517	-0.619	-0.101	-0.479	-0.637	-0.158
ST 2	8	49.2%	-0.698	-0.883	-0.185	-0.815	-0.920	-0.105
ST 3	4	5.7%	-0.913	-0.929	-0.016	-0.948	-0.955	-0.007
ST 3	8	5.8%	-1.190	-1.212	-0.022	-1.509	-1.529	-0.020
ST 4	4	41.5%	-0.659	-0.680	-0.021	-0.958	-0.994	-0.037
ST 4	8	20.6%	-0.665	-0.690	-0.025	-1.045	-1.060	-0.015
ST 5	4	67.2%	-1.282	-1.543	-0.261	-1.487	-1.762	-0.276
ST 5	8	28.8%	-2.107	-2.181	-0.074	-1.782	-1.963	-0.181
ST 6	4	28.7%	-0.356	-0.425	-0.070	-0.483	-0.559	-0.076
ST 6	8	8.7%	-0.361	-0.390	-0.029	-0.461	-0.484	-0.023
ST 7	4	53.9%	-0.564	-0.671	-0.107	-0.741	-0.829	-0.088
ST 7	8	27.4%	-0.745	-0.813	-0.068	-0.897	-0.930	-0.034
ST 8	4	63.9%	-0.054	-0.009	0.045	-0.110	-0.022	0.088
ST 8	8	64.8%	-0.048	0.145	0.192	-0.177	0.025	0.202

Table 8: Post-stratification Weights. This table shows the standardized difference between the average test score in the masked data versus the complete data. For each column in the "Math" and "Reading" panels, the mean in the masked data is computed using post-stratification weights. The "None" column does not involve weights; the "Race/Gender" weights by the joint proportions for race-gender combinations; the "School" column weights by school size; and "All" uses weights according to race, gender, and school.

		Math				Reading			
State	Grade	None	Race/Gender	School	All	None	Race/Gender	School	All
ST 1	4	0.019	0.081	0.013	0.009	0.009	0.092	0.016	0.002
ST 1	8	0.020	-0.005	0.015	0.006	0.019	0.047	0.022	0.012
ST 2	4	0.026	0.201	-0.016	0.039	0.016	0.167	0.000	0.037
ST 2	8	0.023	0.173	0.033	0.077	0.019	0.105	0.042	0.055
ST 3	4	0.002	0.169	-0.112	-0.011	0.001	0.170	-0.111	-0.020
ST 3	8	0.005	0.162	-0.107	-0.015	0.001	0.163	-0.102	-0.030
ST 4	4	0.025	0.132	0.001	0.060	0.031	0.169	-0.011	0.059
ST 4	8	0.007	0.063	0.039	0.082	0.011	0.149	0.021	0.100
ST 5	4	0.037	0.170	-0.057	-0.049	0.035	0.168	-0.041	-0.034
ST 5	8	0.033	0.402	-0.023	0.034	0.035	0.315	0.017	0.055
ST 6	4	0.050	0.056	-0.010	-0.006	0.046	0.106	-0.023	-0.007
ST 6	8	0.025	0.071	0.047	0.045	0.019	0.105	0.024	0.041
ST 7	4	0.045	0.330	-0.017	0.014	0.048	0.330	-0.027	0.009
ST 7	8	0.026	0.353	0.061	0.119	0.025	0.352	0.047	0.110
ST 8	4	0.009	0.060	-0.016	-0.009	0.008	0.046	-0.012	-0.011
ST 8	8	0.007	-0.039	0.002	0.008	0.007	-0.020	-0.001	0.005