Evaluation of Bias Correction Methods for "Worst-case" Selective Non-participation in NAEP

Don McLaughlin
Larry Gallagher
Fran Stancavage
American Institutes for Research

Commissioned by the NAEP Validity Studies (NVS) Panel
May 2004

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

The authors would like to thank Beth Scarloss for her help in preparing this manuscript.

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
1791 Arastradero Road
Palo Alto, CA 94304-1337
Phone: 650/ 493-3550
Fax: 650/ 858-0958

Table of Contents

Introduction **5**

Procedure **5**

Simulation of bias 6
Simulation of corrections 6

Results: School-level Analysis **8**

Simulations of school-level non-participation 8
Simulations of bias correction 10

Results: Student-level Analysis **15**

Simulations of student-level non-participation 15
Simulation of bias correction 16
Correcting for non-participation of students with the lowest two scores 16
Correcting for non-participation of students with the highest two scores 18

Feasibility of Using State Assessment Scores **20**

Summary **21**

Appendix **A-1**

Introduction

With the advent of No Child Left Behind (NCLB), the context for NAEP participation is changing. Whereas in the past participation in NAEP has always been voluntary, participation is now mandatory for some grade and subjects among schools receiving Title I funds. While this will certainly raise school-level participation rates in the mandated grades and subjects, it could have the opposite effect on non-mandated grades and subjects, particularly in light of the increased burden of state testing also required by NCLB. At the student level, participation remains voluntary in all subject areas as it has been in the past. However, participation rates could be influenced negatively by the NCLB requirement for more aggressive notification of students and parents regarding their rights to opt out of NAEP testing.

Random non-participation introduces random error into NAEP estimates. More worrisome is the possibility of selective non-participation at the top or bottom of the ability distribution, which would introduce a bias into statewide mean scores. Although NAEP has not, as had once been proposed, been given an official confirmatory role under NCLB, one can expect greater scrutiny of the relationship between state scores and state NAEP scores in the coming years. This could lead to subtle pressures that depress participation among schools or students near the bottom of the distribution. Conversely, if non-Title I schools decide to take advantage of their exemption from mandatory participation, this could remove a disproportionate number of high performing schools from the sample, given that Title I funds are targeted at schools serving disadvantaged students. Furthermore, participation rates among high performing students could be differentially affected by a variety of factors. These could include a greater reluctance to lose instructional time for testing, or simply a greater willingness among affluent students and parents to assert their rights under the law. It may take several years for these countervailing forces to play themselves out and for any serious problems in NAEP participation rates to manifest. In the meantime, the purpose of this study is to estimate the potential bias from "worst-case" scenarios of selective non-participation, and to examine the extent to which statistical methods can correct for that bias.

Procedure

Since non-participation by schools and non-participation by students each pose distinct questions, we addressed them separately, but with similar analyses. For the simulation study of school-level non-participation bias, we analyzed data from the grade 4 and grade 8 NAEP 2000 mathematics assessment. We began with a set of school-level data for 35 states at grade 4 and 31 states at grade 8 that had both NAEP and state assessment date for that year. The variables of interest were school means for the NAEP and local state assessment, and two school-level demographic variables: 1) percent minority enrollment, and 2) percent of students eligible for free/reduced price lunch. The latter were chosen because of their association with test performance and because they are available uniformly across the states in the Common Core of Data database.

For student-level analyses, we started with grade 4 and grade 8 data from prior studies in which NAEP scores and state assessment scores had been linked for individual students. Linked data were available for 1996 mathematics in 4 states and for 1998 reading in 6 other states.[1] For the current analyses we combined data across subject areas. The variables of interest were individual mean composite plausible values for NAEP, individual state test scores, and individual indicators of gender,

[1] Grade 8 reading data were only available for 5 states.

minority status, disability, and limited English proficiency. As with the school level analyses, the selection of variables was constrained by what was available. In the actual application of the proposed method for correcting for non-participation bias, additional variables might be used.

Simulation of bias

For each state, we successively truncated one tail of the NAEP distribution and recomputed the statewide mean. To simulate biased school-level non-participation, we discarded the 5, 10, 15, 20, and 25 percent of schools with either the lowest or highest mean scores. In order to explore the relationship between bias correction methods and the characteristics of the data set, two different criteria were used to identify schools for deletion. In the first, case truncation was based on actual NAEP scores. In the second case, truncation was based on *predicted* NAEP score, where the predictors were a combination of the aforementioned demographic variables and state assessment scores.

To simulate biased student-level non-participation, we discarded the 1, 2, 3, 5, or 10 students in each school with either (a) the lowest or highest NAEP scores, or (b) the lowest or highest state test scores.

These procedures were repeated over 35 states (grade 4) or 31 states (grade 8), to study school-level non-participation, and over 10 states (grade 4) or 9 states (grade 8), to study student-level nonparticipation. The replications yielded an overall estimate of bias and a standard deviation of that bias across states.

Simulation of corrections

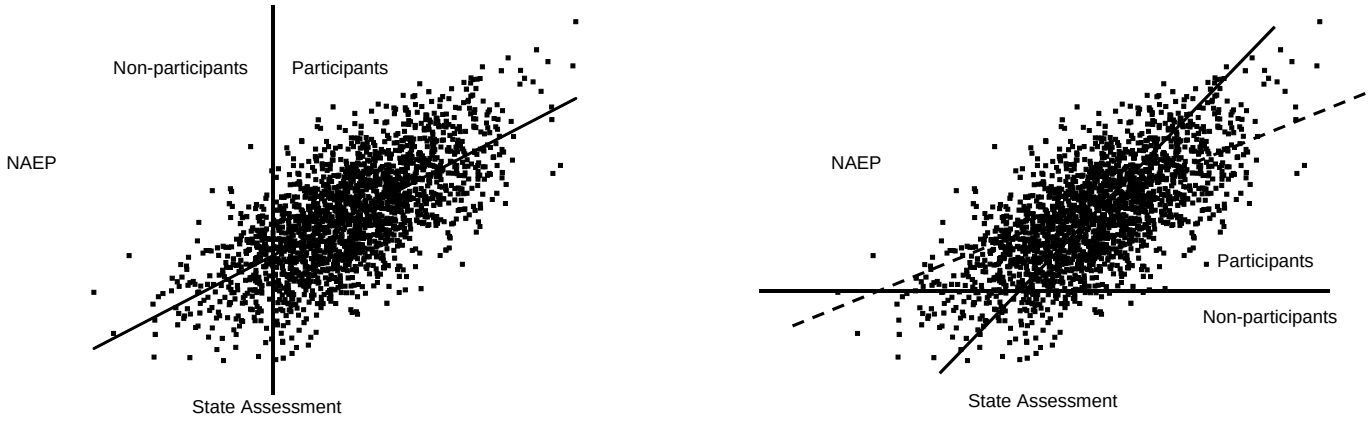
Operationally, NAEP corrects for school non-participation bias by assigning the weights of non-participating schools to demographically similar schools that *did* participate. In this study, we mimicked that correction method by using regression based on demographic predictors to assign means for the "non-participating" schools. Reweighting or regression-based imputation should produce similar results. In separate analyses, we also used demographic predictors to assign scores for non-participating *students*, and we used state assessment scores (alone or in combination with demographic variables) as additional predictors of school means and student scores. After each set of imputations, the state population mean was estimated using the imputed data, and the new mean compared with the statewide mean based on the full NAEP sample.

Because there is concern that linking state assessment scores to NAEP records could possibly be used to identify responses of individual students participating in NAEP (although the likelihood of that event must be considered extremely remote), NAEP is considering the use of categorical scores on state assessments as predictors. To model this, we conducted a second series of imputations in which we replaced the actual values of each of the (continuous) predictors with quartile versions of the predictors. (Demographic predictors at the student level were not affected by this variation since the values for these predictors were dichotomous, e.g., male/female, white/non-white.)

We also tested three different imputation methods in order to compare the effectiveness of the correction when the imputation method matches the omission pattern with the effectiveness of the correction when the imputation method does not match the omission pattern. When there is truncation of cases with a low value on the ***predictor***, as in the left half of Figure 1, standard (forward) linear regression should yield very good estimates of the state-level bias because the regression parameters estimated on the truncated population are expected to be identical to the regression parameters on the full population. However, when there is truncation of cases with low value on the ***dependent variable*** (i.e., NAEP), as in the right half of Figure 1, linear regression will over-predict the omitted scores because the regression parameters based on the truncated population are different from those on the

total population (see the dashed line in Figure 1). This results in an overall under-correction of the bias introduced by the NAEP score omission. A reverse regression (predicting state assessment scores from NAEP scores) provides an accurate basis for estimating the bias in this case. Unfortunately, there is virtually no way to distinguish between circumstances corresponding to the left and right halves of Figure 1 in practice; the actual situation is typically somewhere between these two extremes. Therefore, we have tried both forward and reverse regression, as well as the middle strategy of linear equating, for each truncation scenario.

Figure 1 – Effects of truncation rule on linear regression estimates



In summary, we iterated over the following variations in predictor sets, levels of data categorization, and imputation methods:

1. *Predictors sets*
   - State assessment scores
   - Demographic variables
   - State scores and demographic variables
2. *Levels of data categorization*
   - Continuous values of predictor variables
   - Quartiles of predictor variables
3. *Imputation methods*
   - Forward linear regression
   - Linear equating,[2] where the imputation is based on standardized values of the predictors and regression estimate
   - Reverse regression, where the predictor is regressed on available NAEP scores, and the regression coefficients used to impute the missing NAEP scores

[2] The method known as linear equating adjusts the predictor to have the same mean and variance as the dependent variable. This method is, in an algebraic sense, midway between forward and reverse regression. See Kolen, M. J., & Brennan, R. L. (1995). *Test equating: methods and practices.* New York: Springer-Verlag, pages 30-ff. for further details.

The three imputation methods correspond to:

1) $\hat{y} - \bar{y} = \beta(x - \bar{x}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

2) $\hat{y} - \bar{y} = \frac{1}{r}\beta(x - \bar{x}) = \frac{\sigma_y}{\sigma_x}(x - \bar{x})$

3) $\hat{y} - \bar{y} = \frac{1}{r^2}\beta(x - \bar{x}) = \left(\frac{1}{r}\right)\frac{\sigma_y}{\sigma_x}(x - \bar{x}),$ or $(x - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(\hat{y} - \bar{y})$

where $\hat{y}$ is the NAEP estimate for a non-participant, x is a predictor composite, β is the estimated regression coefficient, and r is the correlation between x and y. β and r are, of course, based on participants.

Results: School-level Analysis

Simulations of school-level non-participation

Figure 2 presents the actual mean bias at grade 4 resulting from truncation at the lower tail of the NAEP score distribution. The graph is nearly a straight line; for every 5 percent of schools removed based on low mean NAEP scores, the average statewide mean estimates (using NAEP weights) are inflated by approximately 1.3 NAEP score points. Correspondingly, the data in Figure 3 show that, for every 5 percent of schools removed based on low *predicted* NAEP scores, the state mean is inflated by approximately 1.1 points. Comparable graphs for grade 8 are given in the appendix; the pattern of bias is very similar to that observed at grade 4.

Figure 2 – Bias in grade 4 state NAEP estimates after truncation of schools with low NAEP scores (average of 35 states, grade 4 math, 2000)

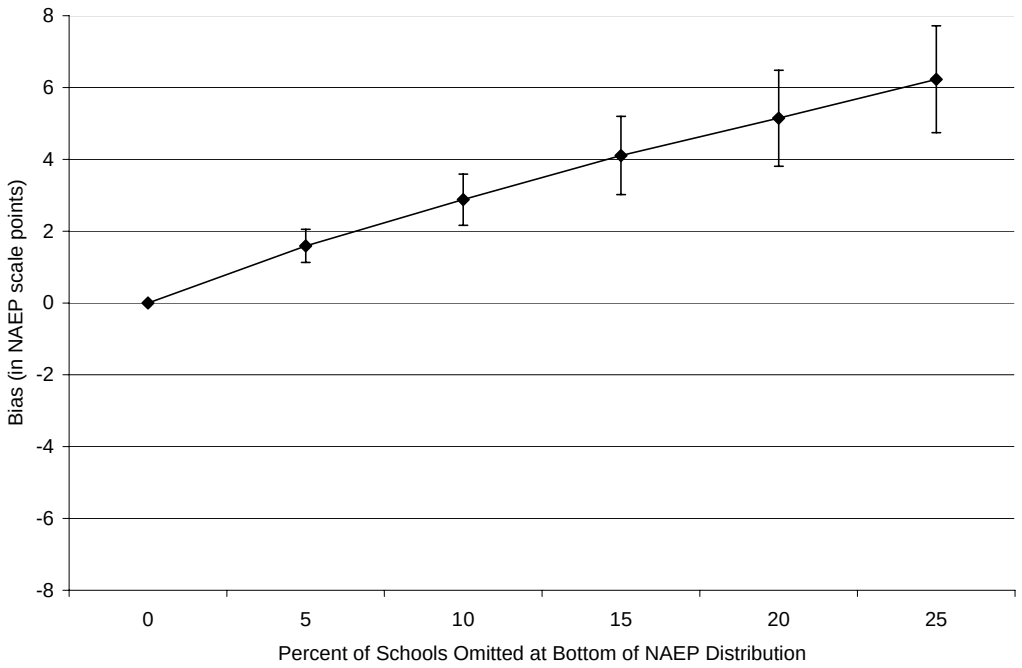
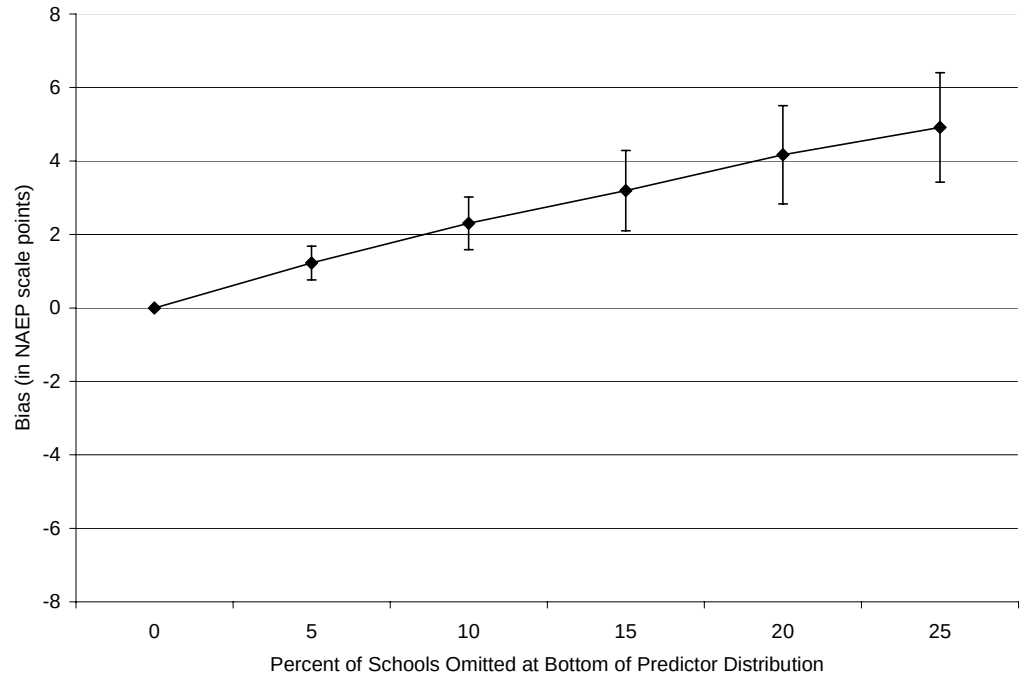


Figure 3 – Bias in grade 4 state NAEP estimates after truncation of schools with low predicted NAEP scores (average of 35 states, grade 4 math, 2000)



It is also worthwhile to note that for both grades the effect is fairly uniform across states. For example, the standard deviation of the bias across states is less than a quarter of the magnitude of the bias in Figure 2 and about a third of the magnitude of the bias in Figure 3.

Simulations of bias correction

Bias corrections were simulated at each level of non-participation. However, because the bias and bias corrections are linear with respect to the percent of schools removed from a tail, this paper focuses on reporting the bias correction results when 10 percent of schools are removed from either tail. Even at this level of non-participation, the bias is large enough to seriously compromise NAEP results unless a correction is applied.[3] In addition, because results are similar at grades 4 and 8, the presentation focuses on grade 4. Grade 8 results are given in the appendix.

Correcting for non-participation of the ten percent of schools with lowest NAEP scores. In Figure 4, we compare the performance of three predictor sets, three imputation methods, and two levels of categorization (continuous data and quartiles) when the bottom 10% of grade 4 schools, based on NAEP scores, are assumed not to have participated.

The mean uncorrected bias is 2.9 NAEP points. When the truncation of the distribution is based on actual NAEP scores, the corrections based on forward regression are only moderately effective. (Recall from the above discussion that regression in which there is truncation of the dependent variable is expected to result in under-correction of the bias.) Imputing these missing scores by regressing on either demographics or statewide test scores alone reduces the bias to between 1.6 to 1.8 NAEP points, while combining the two predictors reduces the bias to 1.4 NAEP points. Note that the use of predictor quartiles only slightly decreases the effectiveness of the regression imputation.

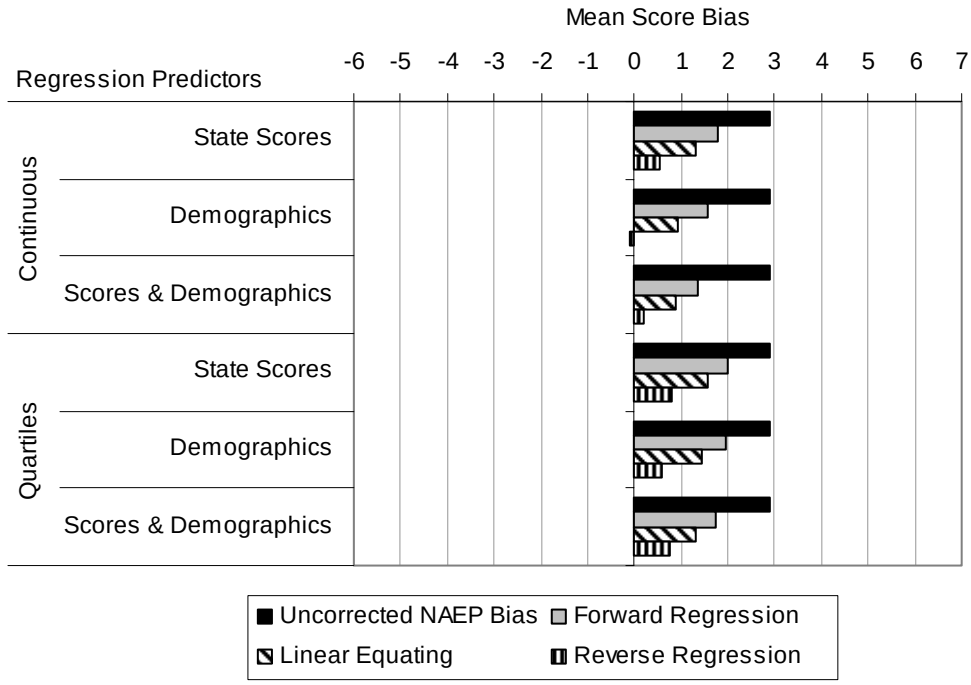
As expected when the truncation is based on the dependent variable, reverse regression imputation performs admirably, reducing the bias to less than 0.5 NAEP points with states scores as a predictor, and to a magnitude of less than 0.2 NAEP points when demographic data is combined with state scores.[4] Even when quartiles of predictors are used, reverse regression still reduces the bias to approximately 0.7 NAEP points.

The method of linear equating is more effective than forward regression, but less effective than reverse regression. The bias is reduced to between 0.9 to 1.3 NAEP points with continuous predictors, and 1.3 to 1.6 points with quartile predictors.

[3] At grade 4, bias is 2.9 NAEP scale points when low score truncation is based on actual NAEP scores, and 2.3 points when truncation is based on predicted NAEP scores. At the top end of the scale, truncation based on actual NAEP scores produces a bias of -2.8 NAEP scale points, while truncation based on predicted NAEP scores yields a bias of -2.3 points.

[4] To impute using reverse regression with a multivariate predictor, we first created the linear composite that best predicts the NAEP score and then applied reverse regression on that linear composite.

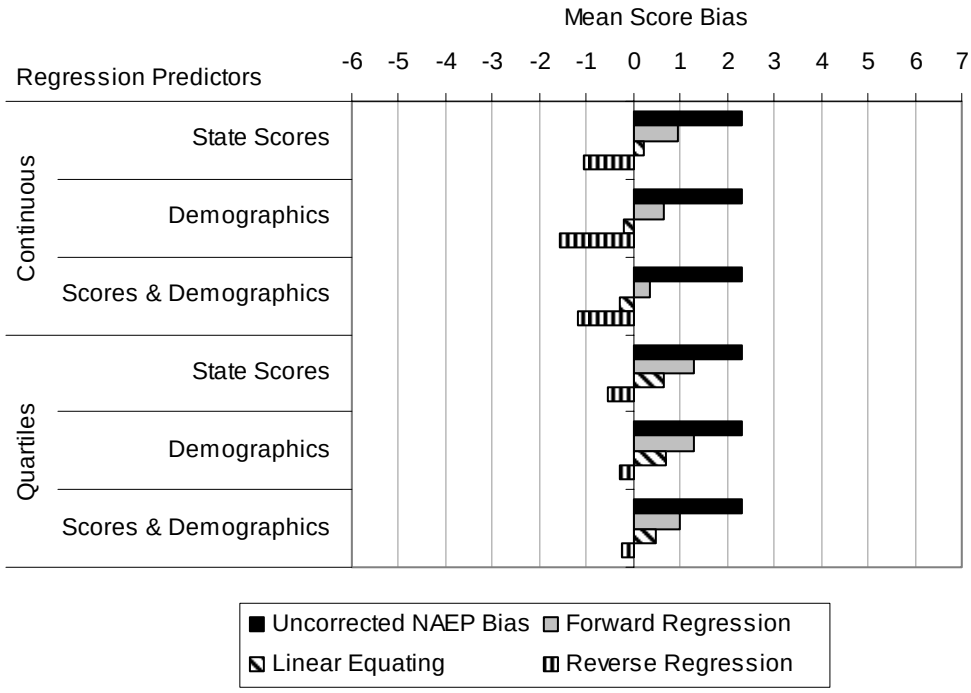
Figure 4 – Grade 4 bias and bias corrections for non-participation of schools in the bottom 10% of NAEP scores, by predictor set and imputation method



Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	2.9	(0.6)	1.8	(0.7)	1.3	(0.7)	0.5	(0.8)
	Demographics	2.9	(0.6)	1.6	(0.4)	0.9	(0.6)	-0.1	(1.0)
	Scores & Demographics	2.9	(0.6)	1.4	(0.4)	0.9	(0.4)	0.2	(0.6)
Quartiles of Predictors	State Scores	2.9	(0.6)	2.0	(0.6)	1.6	(0.6)	0.8	(0.8)
	Demographics	2.9	(0.6)	2.0	(0.4)	1.4	(0.4)	0.6	(0.5)
	Scores & Demographics	2.9	(0.6)	1.7	(0.5)	1.3	(0.4)	0.8	(0.5)

Correcting for non-participation of the 10 percent of schools with lowest predicted NAEP scores. When the simulated non-participation is based on predicted NAEP scores, we observe improved performance of the forward regression method (see Figure 5). However, reverse regression results in significant *over-correction* of the bias, by more than 1 NAEP point, when using continuous predictors. The method of linear equating gives the best overall performance of bias reduction, particularly when state test scores are the only predictor.

Figure 5 – Grade 4 bias and bias corrections for non-participation of schools in the bottom 10% of predicted NAEP scores, by predictor set and imputation method



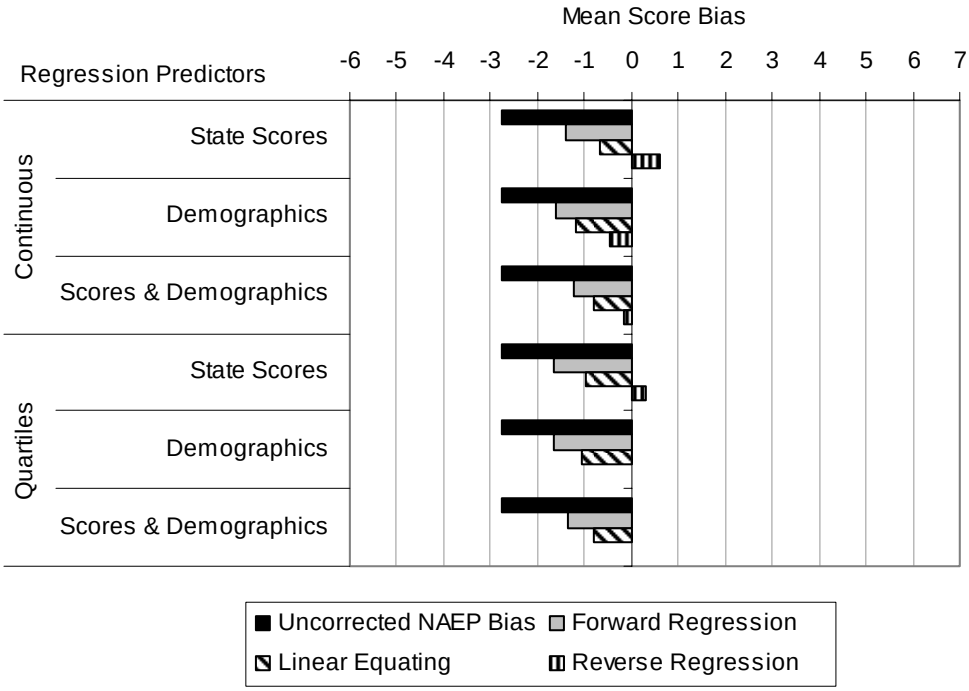
Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	2.3	(0.7)	0.9	(0.8)	0.2	(0.9)	-1.1	(1.4)
	Demographics	2.3	(0.7)	0.6	(0.6)	0.2	(0.6)	-1.6	(1.3)
	Scores & Demographics	2.3	(0.7)	0.3	(0.6)	0.3	(0.6)	-1.2	(0.7)
Quartiles of Predictors	State Scores	2.3	(0.7)	1.3	(0.7)	0.6	(0.6)	-0.5	(1.3)
	Demographics	2.3	(0.7)	1.3	(0.6)	0.7	(0.7)	-0.3	(0.7)
	Scores & Demographics	2.3	(0.7)	1.0	(0.6)	0.5	(0.8)	-0.2	(0.8)

Thus, both forward and reverse regressions are imperfect (when r^2 is less than 1.0): in the one case there is a risk of under-correcting of bias, and in the other case there is a risk of over-correcting, depending upon the underlying (and unknowable) characteristics of the data. That is, because one cannot know, in general, the extent to which school non-participation at the low end of the distribution is more closely correlated with the measured predictors of NAEP scores or with other (unmeasured) characteristics that would more closely match actual NAEP scores, it is impossible in practice to choose between forward and reverse regression on technical grounds. Conceivably, for policy reasons, one might select the method that potentially over-corrects, because it essentially adds

a penalty for non-participation, while a method that under-corrects would favor states with selective school non-participation.

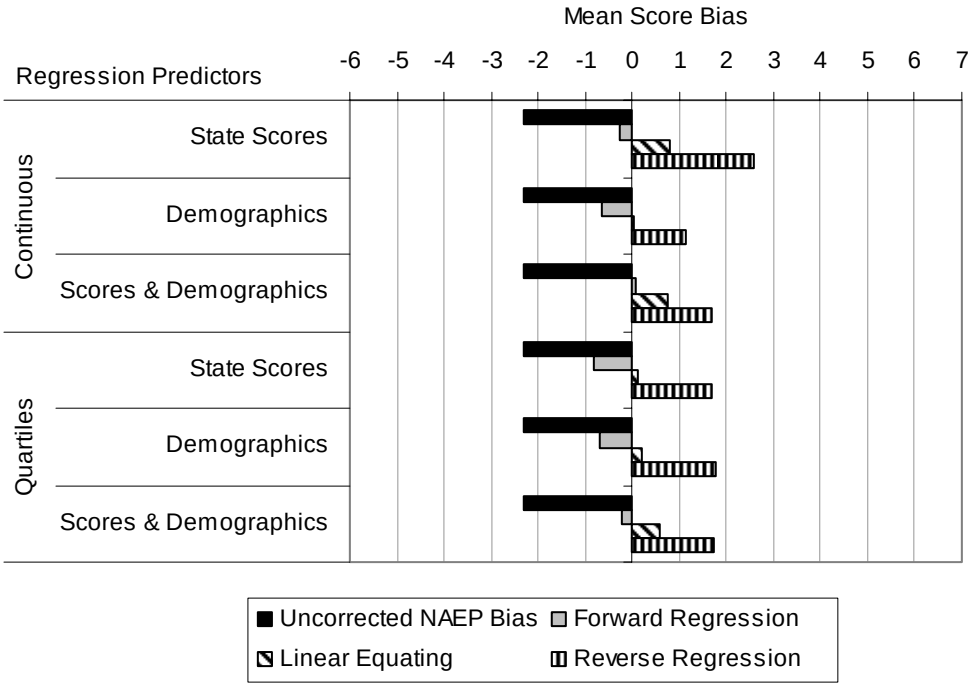
Correcting for selective non-participation by high performing schools. Before turning to a simulation of student-level selective non-participation bias, we considered the potential problem of non-participation of schools with the *highest* NAEP scores or highest predicted NAEP scores. On average, our findings were as expected. As shown in Figure 6 (non-participation by the top ten percent of schools by NAEP score) and Figure 7 (non-participation by the of top ten percent of schools by predicted NAEP score), the results were roughly symmetric.

Figure 6 – Grade 4 bias and bias corrections for non-participation of schools in the top 10% of NAEP scores, by predictor set and imputation method



Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	-2.8	(0.4)	-1.4	(0.5)	-0.7	(0.5)	0.6	(1.0)
	Demographics	-2.8	(0.4)	-1.6	(0.3)	-1.2	(0.3)	-0.4	(0.8)
	Scores & Demographics	-2.8	(0.4)	-1.2	(0.4)	-0.8	(0.4)	-0.2	(0.6)
Quartiles of Predictors	State Scores	-2.8	(0.4)	-1.6	(0.4)	-1.0	(0.4)	0.3	(0.7)
	Demographics	-2.8	(0.4)	-1.7	(0.3)	-1.1	(0.4)	0.0	(0.8)
	Scores & Demographics	-2.8	(0.4)	-1.3	(0.3)	-0.8	(0.3)	0.0	(0.6)

Figure 7 – Grade 4 bias and bias corrections for non-participation of schools in the top 10% of predicted NAEP scores, by predictor set and imputation method



Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	-2.3	(0.6)	-0.2	(1.1)	0.8	(0.9)	3.2	(3.3)
	Demographics	-2.3	(0.6)	-0.7	(1.3)	0.0	(0.7)	3.3	(14.2)
	Scores & Demographics	-2.3	(0.6)	0.1	(1.2)	0.8	(0.7)	3.0	(8.6)
Quartiles of Predictors	State Scores	-2.3	(0.6)	-0.8	(0.8)	0.21	(0.9)	1.7	(1.2)
	Demographics	-2.3	(0.6)	-0.7	(1.0)	0.2	(0.8)	1.8	(1.5)
	Scores & Demographics	-2.3	(0.6)	-0.2	(0.8)	0.6	(0.8)	1.7	(1.2)

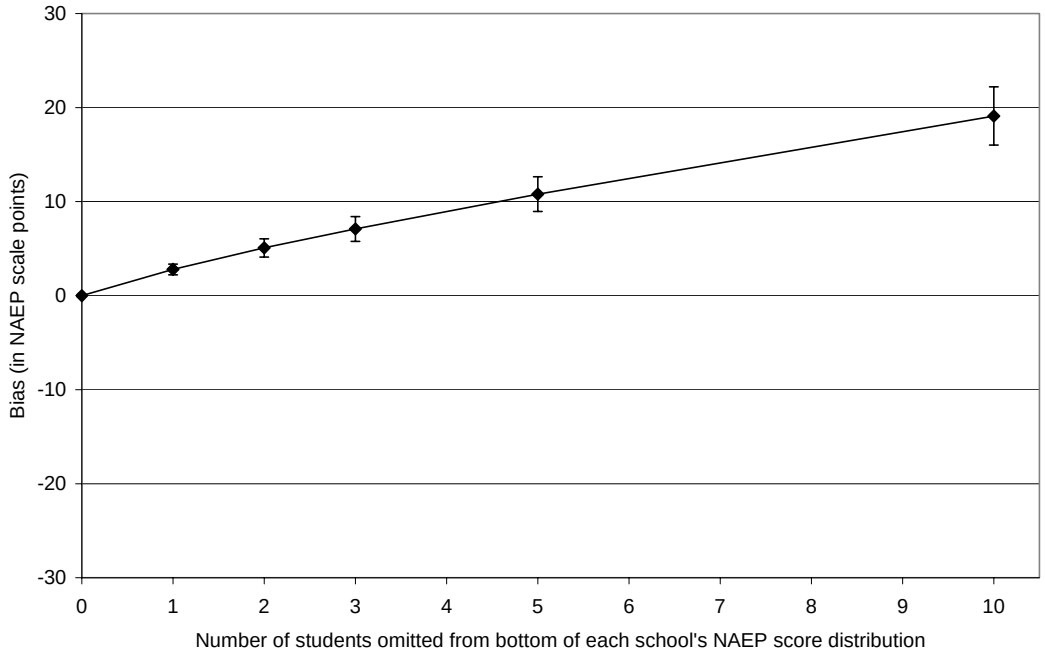
The policy considerations affecting the choice of imputation method are different at the two tails of the distribution. In order to avoid rewarding states with selective non-participation at the top of the distribution, one must avoid any method that over-corrects the bias (i.e., any method that inflates a state mean based on imputation above the value that would have been obtained if all schools had participated in NAEP). From this perspective, the reverse regression method is most problematic, as it can lead to substantial overcorrection in the case where non-participation is most closely correlated with the measured predictors of NAEP scores.

Results: Student-level Analysis

Simulations of student-level non-participation

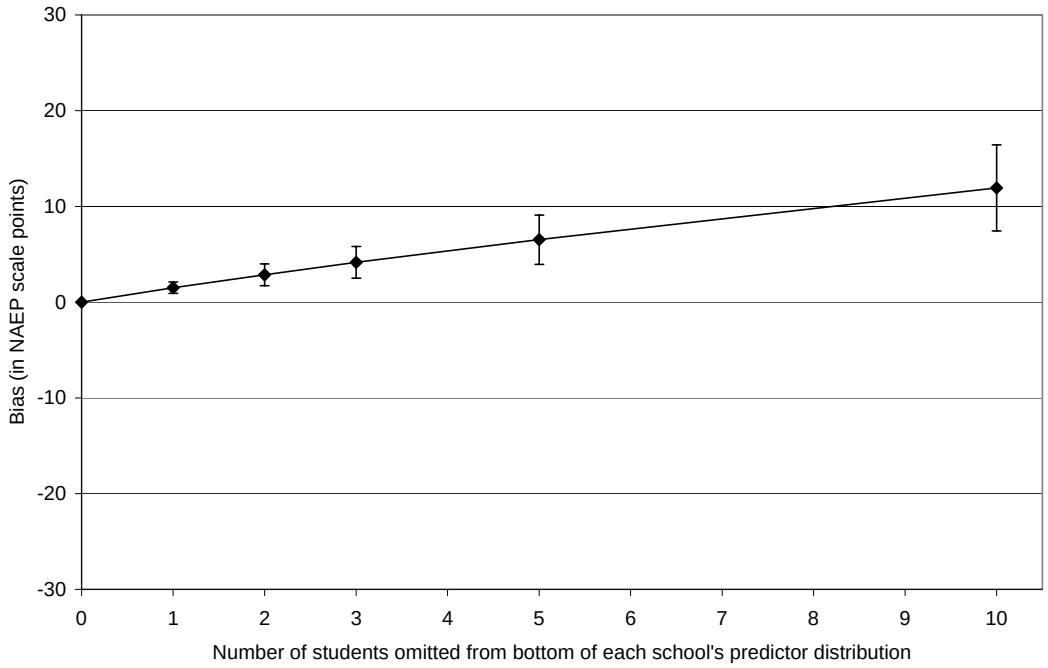
Figure 8 presents the actual mean bias at grade 4 resulting from lower-tail NAEP score truncation of students. We see that for every student removed from the lower tail of the distribution in each school, based on NAEP scores, the average statewide mean estimates (using NAEP weights) are inflated by approximately 2.5 NAEP score points.[5] In the corresponding graph of bias resulting from truncation based on state assessment scores (Figure 9), the state mean at grade 4 is inflated by approximately 1.5 points for every student-per-school removed. Thus, particularly in the case of truncation based on NAEP scores, the size of the bias is substantial and considerably greater than the bias from selective *school* non-participation. The loss of just the two lowest scoring students from each school, if not corrected, would seriously compromise NAEP results. Comparable graphs for grade 8 are given in the appendix.

Figure 8 - Bias in grade 4 state NAEP estimates after truncation of students with low NAEP scores (average of 10 states, NAEP reading, 1998 or mathematics, 1996)



[5] The graph is slightly non-linear. This pattern occurs because the average number of students per session is typically only slightly more than 20; removing the lowest 10 students per session removes some students from the middle of the distribution, and removing students in the middle of the distribution does not contribute to bias.

Figure 9 – Bias in grade 4 state NAEP estimates after truncation of students with low state assessment scores (average of 10 states, state assessment scores in reading or mathematics)



As with the school-level truncation, the effect is fairly uniform across states. The standard deviation of the bias across states is less than one-fifth of the magnitude of the bias in Figure 8 and about two-fifths of the magnitude of the bias in Figure 9.

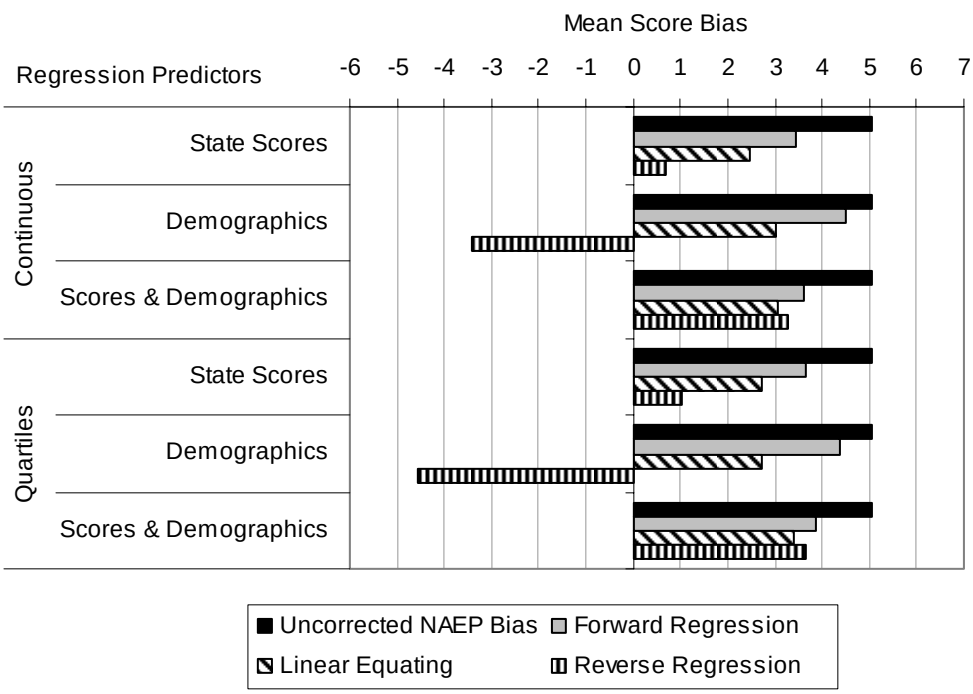
Simulation of bias correction

The following discussion centers on grade 4 bias correction results when 2 students are removed from the either tail of the distribution of students within schools. Grade 8 results are given in the appendix.

Correcting for non-participation of students with the lowest two scores

When the two lowest performing students (by NAEP score) are dropped from each school, the mean NAEP score bias is 5.1 points at grade 4 (see Figure 10). Imputing scores via forward regression reduces this bias to 3.1 points when both state test scores and demographics are used as predictors, while use of linear equating reduces the bias to 2.0 points, about 40 percent of the original bias value. Reverse regression produces an over-correction, particularly when demographic factors alone are used as predictors.

Figure 10 – Grade 4 bias and bias corrections for non-participation of the 2 students in each school with the lowest NAEP scores, by predictor set and imputation method

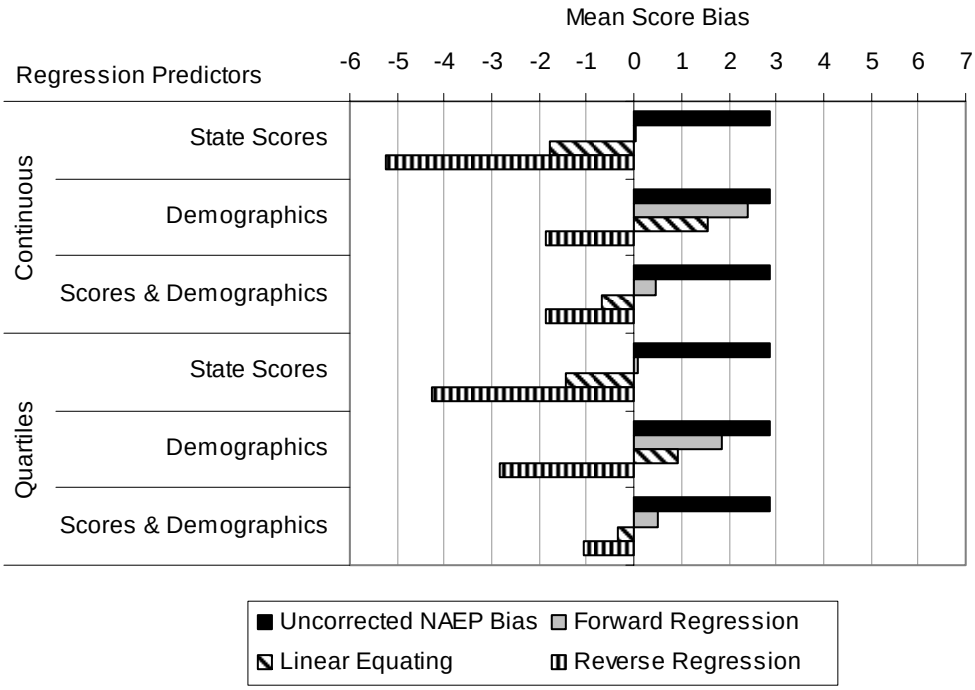


Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	5.1	(1.0)	3.4	(1.0)	2.4	(1.1)	0.7	(1.4)
	Demographics	5.1	(1.0)	4.5	(0.8)	3.2	(0.8)	-1.8	(1.5)
	Scores & Demographics	5.1	(1.0)	3.1	(0.9)	2.0	(0.7)	-0.2	(0.8)
Quartiles of Predictors	State Scores	5.1	(1.0)	3.7	(0.9)	2.7	(0.9)	1.0	(1.3)
	Demographics	5.1	(1.0)	4.5	(0.8)	3.2	(0.8)	-1.8	(1.5)
	Scores & Demographics	5.1	(1.0)	3.4	(0.9)	2.3	(0.7)	0.1	(0.9)

In the student simulations, we are interested particularly in the impact of replacing the continuous state score with a quartile state score in the imputations. This is because concerns about privacy arise in the case of individual student scores, rather than mean school scores. In fact, we see that the substitution of quartile scores does not seriously degrade the corrections. For example, in the case of the linear regression, the residual bias only increases from 2.0 to 2.3 with the substitution of quartiles.

When the two lowest performing students (by state score) are dropped from each school, the mean NAEP score bias is 2.9 points (Figure 11). As expected when truncating the data by the predictor, forward regression reduces the bias to practically 0 when state scores and demographics are used as predictors. The methods of linear equating and reverse regression, however, resulted in significant over-correction in most cases.

Figure 11 – Grade 4 bias and bias corrections for non-participation of the 2 students in each school with the lowest state assessment scores, by predictor set and imputation method



Mean Bias (Standard Deviation)		Uncorrected NAEP Bias	Forward Regression	Linear Equating	Reverse Regression
Continuous Predictors	State Scores	2.9 (1.1)	0.1 (0.3)	-1.8 (1.0)	-5.2 (3.2)
	Demographics	2.9 (1.1)	2.4 (1.0)	1.6 (0.6)	-1.0 (1.3)
	Scores & Demographics	2.9 (1.1)	0.0 (0.3)	-1.5 (0.9)	-4.0 (2.6)
Quartiles of Predictors	State Scores	2.9 (1.1)	0.7 (0.3)	-0.9 (0.7)	-3.9 (2.8)
	Demographics	2.9 (1.1)	2.4 (1.0)	1.6 (0.6)	-1.0 (1.3)
	Scores & Demographics	2.9 (1.1)	0.6 (0.3)	-0.7 (0.5)	-2.9 (1.9)

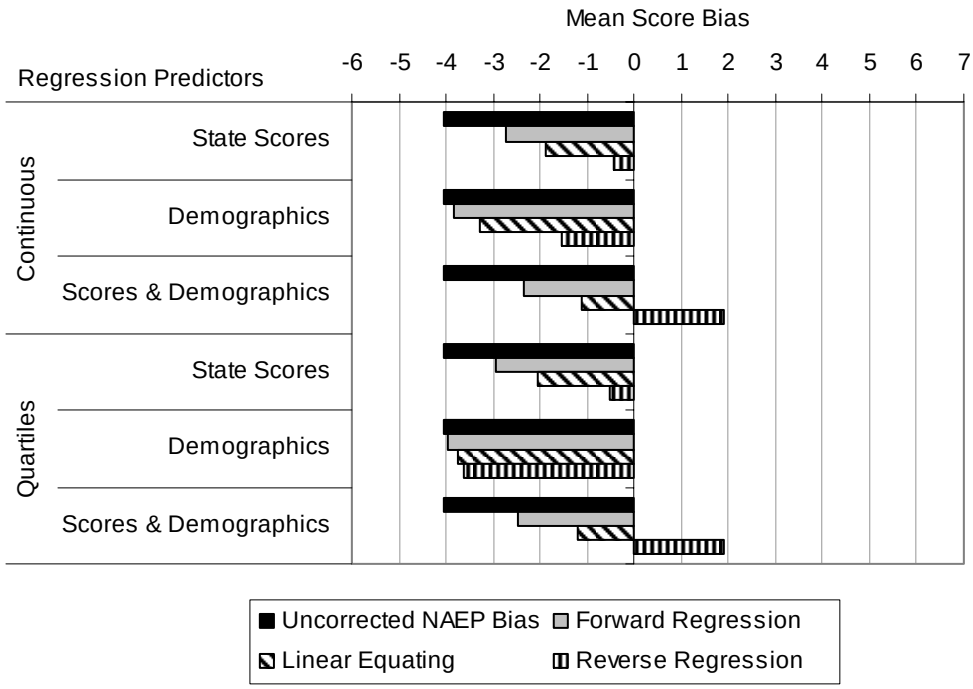
Given that we cannot know the exact mechanism of non-participation of the lower tail of the score distribution, forward regression is indicated as the method that best avoids over-correction. Both linear equating and reverse regression run the risk of significant over-correction.

Correcting for non-participation of students with the highest two scores

As in the case of selective non-participation by high performing schools, our results in the student score bias were generally as expected. As shown in Figure 12 (omission by NAEP score) and Figure 13 (omission by state score), the results were roughly symmetric. When truncation is based on NAEP scores, forward regression is not very successful in removing the bias in this set of simulations. (In

the best case, bias is reduced from -4.0 to -2.6.) When truncation is based on state scores, however, both reverse and linear regression can produce significant over-corrections.[6]

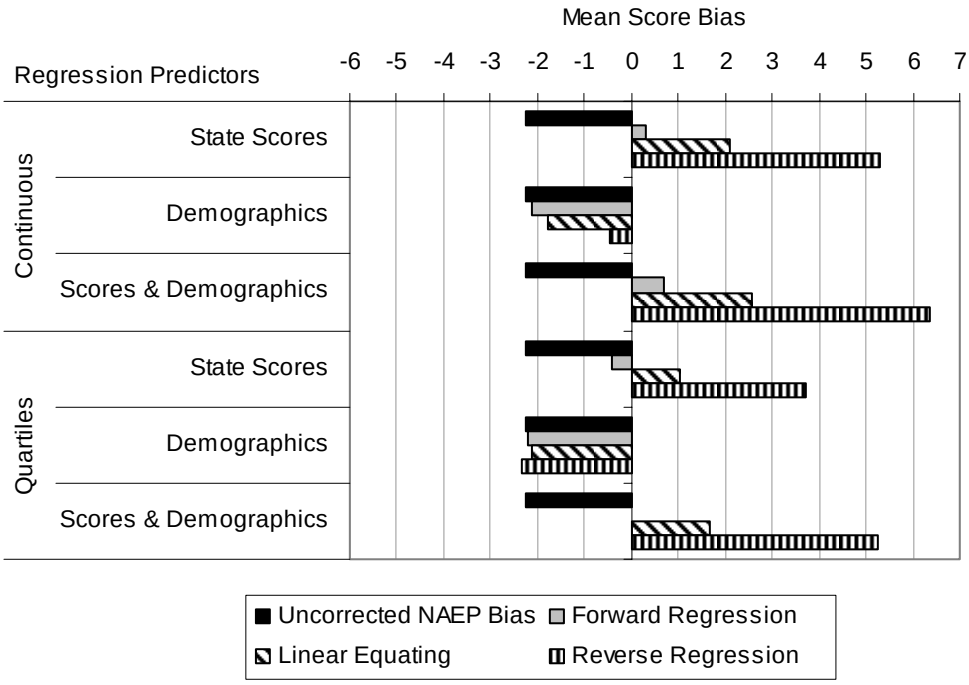
Figure 12 – Grade 4 bias and bias corrections for non-participation of the 2 students in each school with the highest NAEP scores, by predictor set and imputation method



Mean Bias (Standard Deviation)		Uncorrected NAEP Bias	Forward Regression	Linear Equating	Reverse Regression
Continuous Predictors	State Scores	-4.0 (0.6)	-2.7 (0.8)	-1.9 (0.9)	-0.4 (1.3)
	Demographics	-4.0 (0.6)	-3.9 (0.6)	-3.3 (0.5)	-1.6 (0.5)
	Scores & Demographics	-4.0 (0.6)	-2.6 (0.7)	-1.8 (0.7)	-0.4 (0.4)
Quartiles of Predictors	State Scores	-4.0 (0.6)	-2.9 (0.7)	-2.1 (0.8)	-0.5 (1.2)
	Demographics	-4.0 (0.6)	-3.9 (0.6)	-3.3 (0.5)	-1.6 (0.5)
	Scores & Demographics	-4.0 (0.6)	-2.8 (0.7)	-2.0 (0.6)	-0.5 (0.3)

[6] Since truncation in the student-level simulations was based on state scores alone, imputations based only on demographics do not exhibit the same tendency to over-correction. However, in the real world, non-participation could be correlated with demographics as well as state scores.

Figure 13 – Grade 4 bias and bias corrections for non-participation of the 2 students in each school with the highest state assessment scores, by predictor set and imputation method



Mean Bias (Standard Deviation)		Uncorrected NAEP Bias	Forward Regression	Linear Equating	Reverse Regression
Continuous Predictors	State Scores	-2.2 (1.0)	0.3 (0.3)	2.1 (1.3)	5.3 (4.1)
	Demographics	-2.2 (1.0)	-2.1 (0.9)	-1.7 (0.9)	-0.5 (1.1)
	Scores & Demographics	-2.2 (1.0)	0.2 (0.3)	1.6 (0.8)	3.9 (1.8)
Quartiles of Predictors	State Scores	-2.2 (1.0)	-0.4 (0.3)	1.0 (0.7)	3.7 (2.5)
	Demographics	-2.2 (1.0)	-2.2 (1.0)	-2.1 (1.3)	-0.5 (1.1)
	Scores & Demographics	-2.2 (1.0)	0.0 (0.4)	1.7 (1.0)	2.7 (1.5)

Feasibility of Using State Assessment Scores

States will be reluctant to share individual student level data unless these data are free of identifiers. To address this issue, the merging of state scores with NAEP identifiers could be done within each state Department of Education, perhaps at the same time that NAEP samples are being drawn.

This procedure should be field tested in at least three states, varied by their control of state level assessment data. For example, some states depend on intermediary units (large cities, regional centers) for their assessment files; others depend on large contracted companies, while others have mixtures of centralized and decentralized processes. It would be valuable to evaluate the procedures involved in preparing data files within states for eventual merge with NAEP data and to draw conclusions on the readiness of the states and the accuracy of the information.

Summary

At the school level, truncation of the lowest 10 percent of schools in the grade 4 sample (based on NAEP scores) yields an average bias of 2.9 points, ± 0.6 points, across 35 states. At the student level, truncation of the lowest 10 percent in each school (i.e., 2 students) yields a larger bias on average: 5.1 points, ±1 point, across 10 states. In both cases, the bias, which increases roughly in proportion to the percentage of the sample truncated, is remarkably similar across the various states.

If the truncation is based on predicted NAEP scores (school) or state test scores (student), rather than NAEP scores, the bias is less. At the school level, it is about 20 percent less (i.e., 2.3 ±0.7); and at the student level it is reduced by nearly half (i.e., 2.9 ±1.1).

The results for grade 8 were very similar. Specifically, the truncation of the lowest 10 percent of schools in the grade 8 sample (based on NAEP scores) creates an average bias of 3.0 points, ±1.0, across 31 states. At the student level, deletion of the two students with the lowest NAEP scores in each school yields an average bias of 4.8 points, ±0.8, across 9 states.

Again as at grade 4, the bias is less if the truncation is based on predicted NAEP scores or state test scores rather than NAEP scores. At the school level, bias in the grade 8 sample is 2.5, ±1.0, and at the student level it is 3.0, ±0.8.

Correction for non-participation bias is partially effective, eliminating roughly half of the bias when the linear equating method is used. Other regression models can improve the corrections, but their accuracy is dependent on knowledge about the mechanisms of non-participation, which is not likely to be available in practice. Using the reverse regression method can yield over-corrections. That is, the resulting "corrected" state mean estimates for non-participation by low performers can be lower than the results would have been if there had been full participation, while the "corrected" state mean estimates for non-participation by high performers can be higher than the results would have been if there had been full participation.

A useful finding is that state test scores can be replaced with quartiles in these non-participation bias corrections without serious degradation of results. This is important because quartile scores, unlike exact test scores, cannot be used to identify individual NAEP participants. NAEP can thus work with state education agencies to create categorical transformations of state test scores that cannot breach confidentiality or be used to link NAEP responses to individual students.

Appendix

Figure A1 – Bias in grade 8 state NAEP estimates after truncation of schools with low NAEP scores (average of 31 states, grade 8 math, 2000)

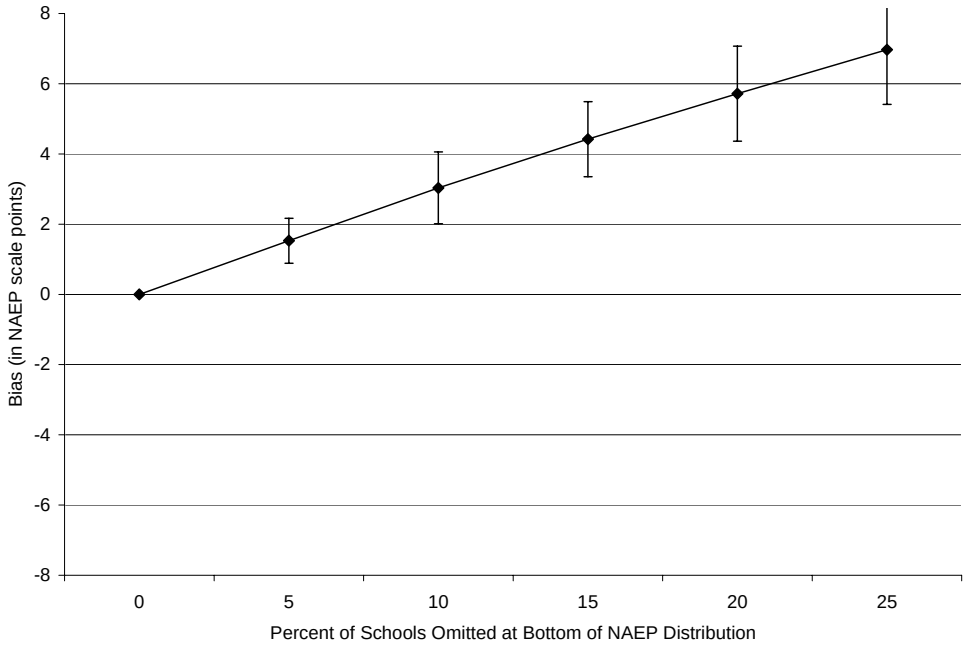


Figure A2 – Bias in grade 8 state NAEP estimates after truncation of schools with low predicted NAEP scores (average of 31 states, grade 8 math, 2000)

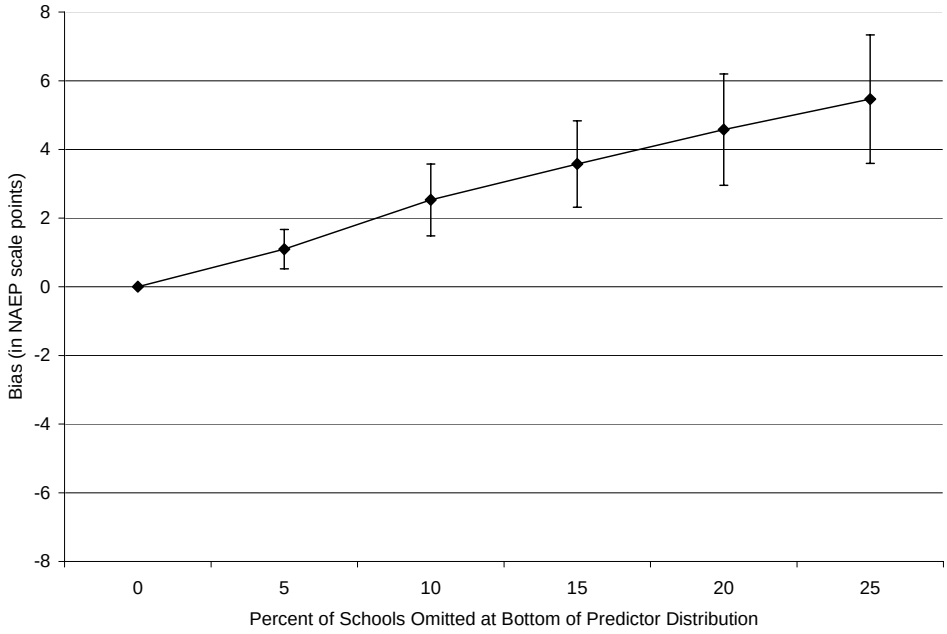


Figure A3 - Bias in grade 8 state NAEP estimates after truncation of students with low NAEP scores (average of 9 states, NAEP reading, 1998 or mathematics, 1996)

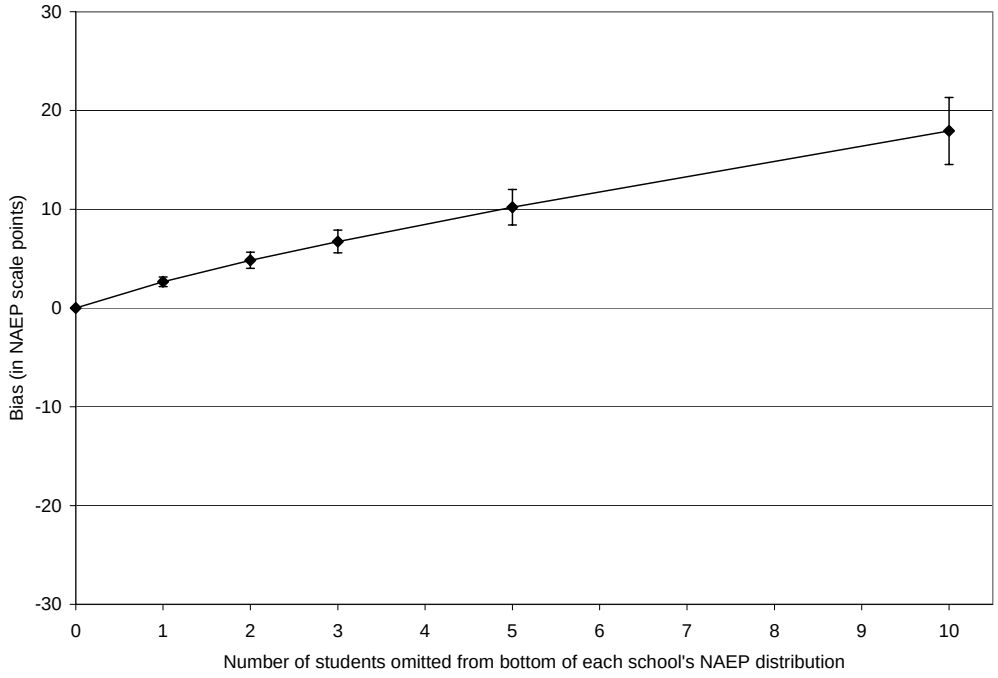


Figure A4 – Bias in grade 8 state NAEP estimates after truncation of students with low state assessment scores (average of 9 states, state assessment scores in reading or mathematics)

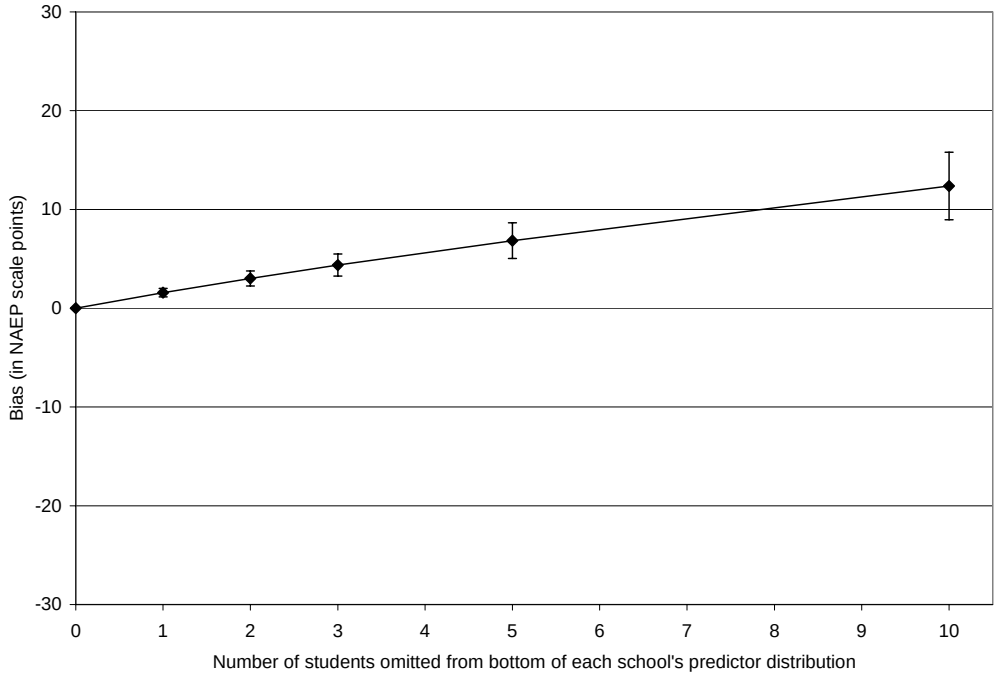
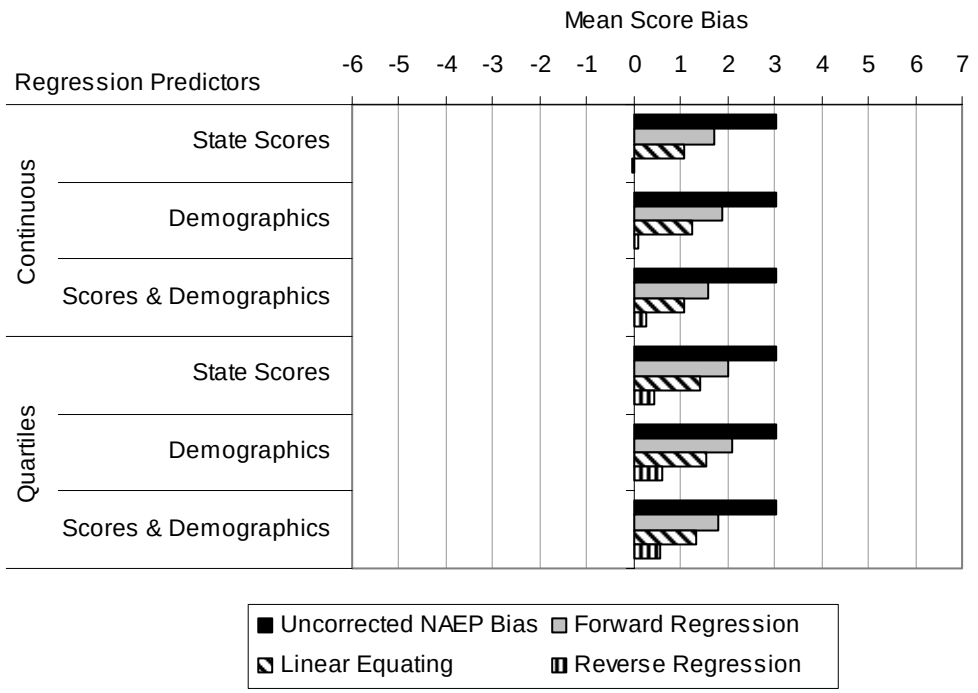
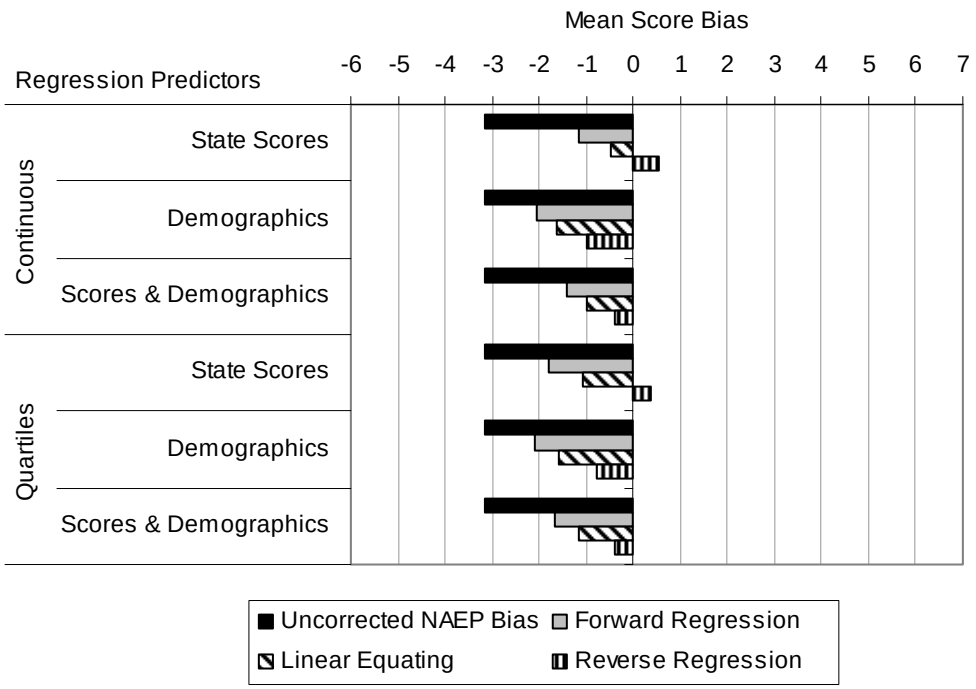


Figure A5 – Grade 8 bias and bias corrections for non-participation of schools in the bottom 10% of NAEP scores, by predictor set and imputation method



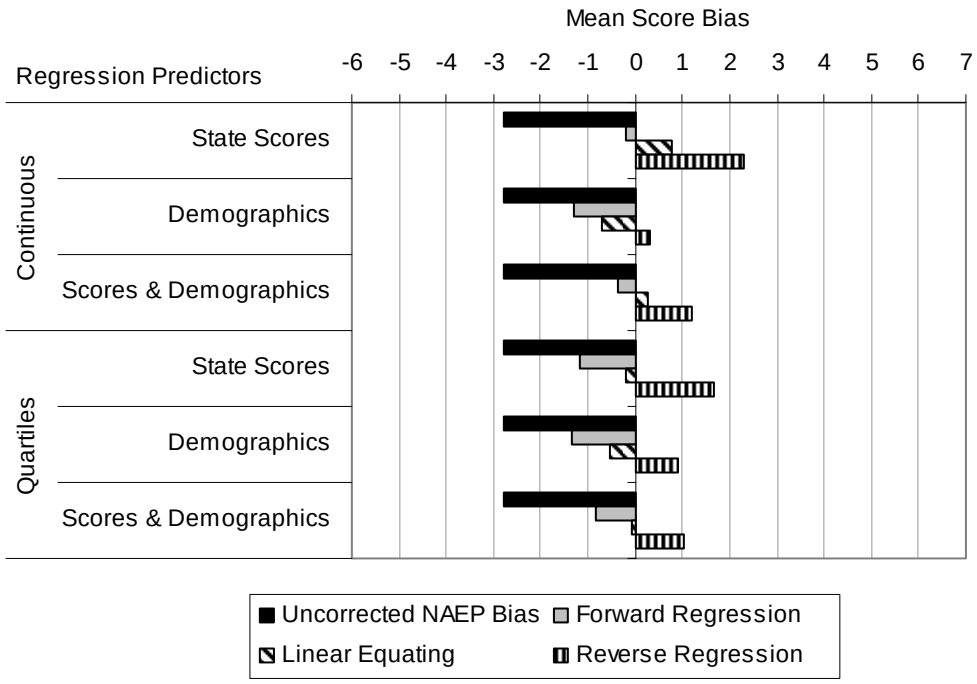
Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	3.0	(1.0)	1.7	(0.7)	1.1	(0.8)	-0.0	(1.4)
	Demographics	3.0	(1.0)	1.9	(0.8)	1.3	(1.1)	0.1	(2.0)
	Scores & Demographics	3.0	(1.0)	1.6	(0.7)	1.1	(0.8)	0.3	(1.3)
Quartiles of Predictors	State Scores	3.0	(1.0)	2.0	(0.7)	1.4	(0.6)	0.5	(0.8)
	Demographics	3.0	(1.0)	2.1	(0.7)	1.6	(0.7)	0.6	(1.0)
	Scores & Demographics	3.0	(1.0)	1.8	(0.7)	1.3	(0.6)	0.6	(0.8)

Figure A6 – Grade 8 bias and bias corrections for non-participation of schools in the bottom 10% of predicted NAEP scores, by predictor set and imputation method



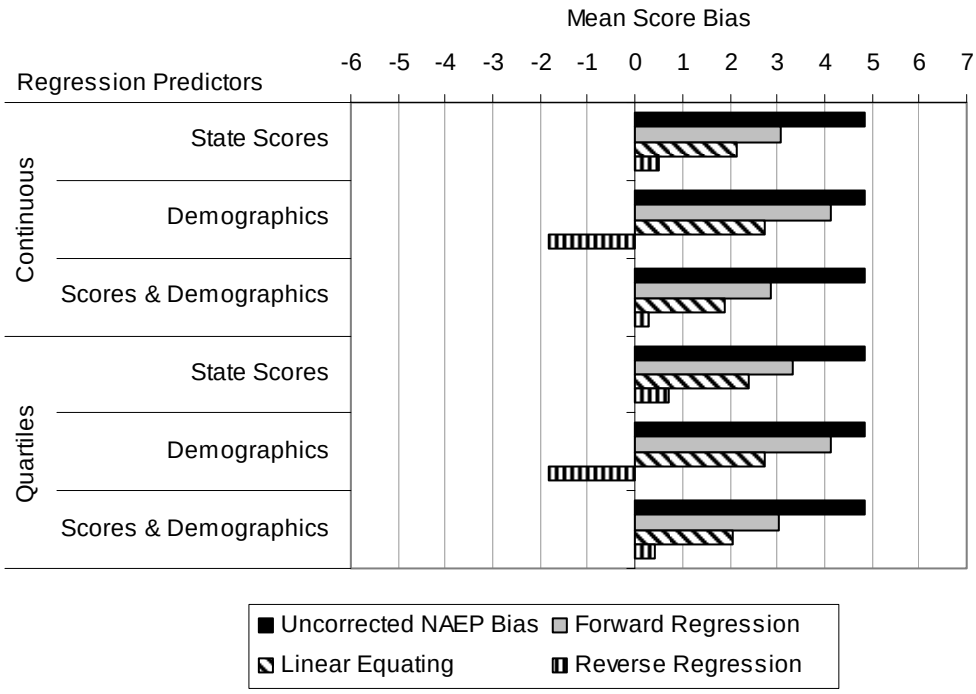
Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	2.5	(1.0)	0.8	(0.9)	0.1	(1.0)	-1.3	(1.7)
	Demographics	2.5	(1.0)	1.0	(1.0)	-0.0	(1.4)	-1.9	(2.7)
	Scores & Demographics	2.5	(1.0)	0.5	(1.0)	-0.2	(1.1)	-1.2	(1.6)
Quartiles of Predictors	State Scores	2.5	(1.0)	1.3	(0.9)	0.6	(0.9)	-0.4	(1.1)
	Demographics	2.5	(1.0)	1.3	(0.8)	0.6	(0.8)	-0.9	(1.3)
	Scores & Demographics	2.5	(1.0)	1.0	(0.8)	0.4	(0.9)	-0.6	(1.2)

Figure A7 – Grade 8 bias and bias corrections for non-participation of schools in the top 10% of NAEP scores, by predictor set and imputation method



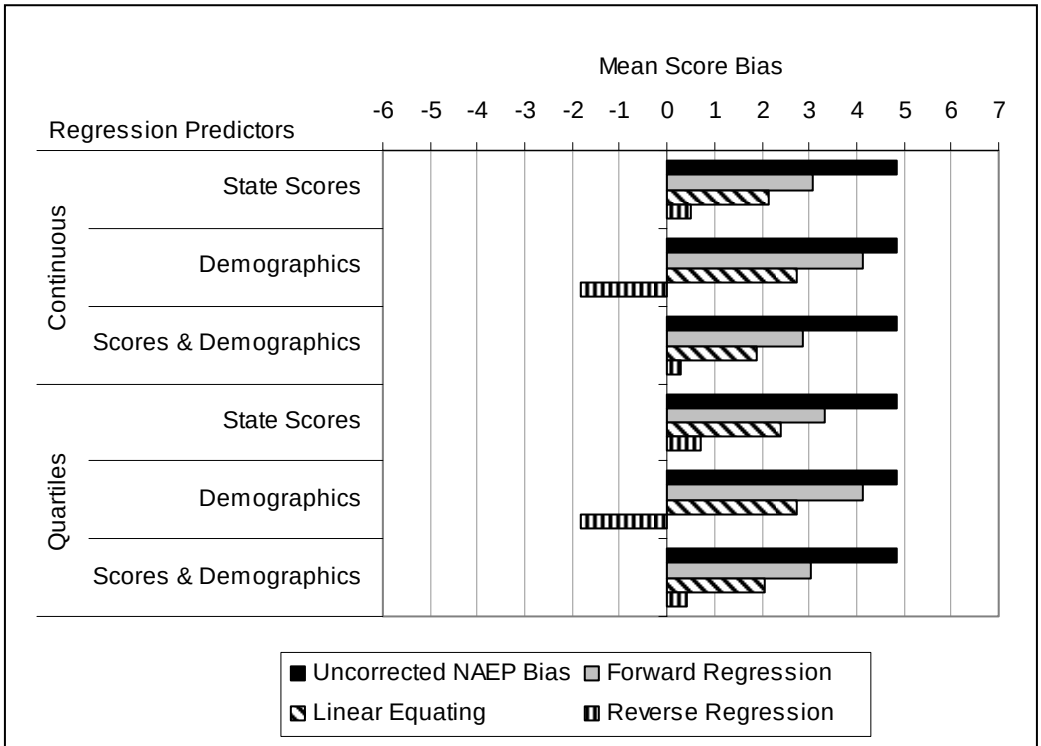
Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	-3.2	(1.1)	-1.2	(0.7)	-0.5	(0.9)	0.6	(1.4)
	Demographics	-3.2	(1.1)	-2.0	(0.7)	-1.6	(0.8)	-1.0	(1.0)
	Scores & Demographics	-3.2	(1.1)	-1.4	(0.8)	-1.0	(0.9)	-0.4	(1.1)
Quartiles of Predictors	State Scores	-3.2	(1.1)	-1.8	(0.8)	-1.1	(0.8)	0.4	(2.7)
	Demographics	-3.2	(1.1)	-2.1	(0.8)	-1.6	(0.8)	-0.8	(1.1)
	Scores & Demographics	-3.2	(1.1)	-1.7	(0.8)	-1.2	(0.9)	-0.4	(1.3)

Figure A8 – Grade 8 bias and bias corrections for non-participation of schools in the top 10% of predicted NAEP scores, by predictor set and imputation method



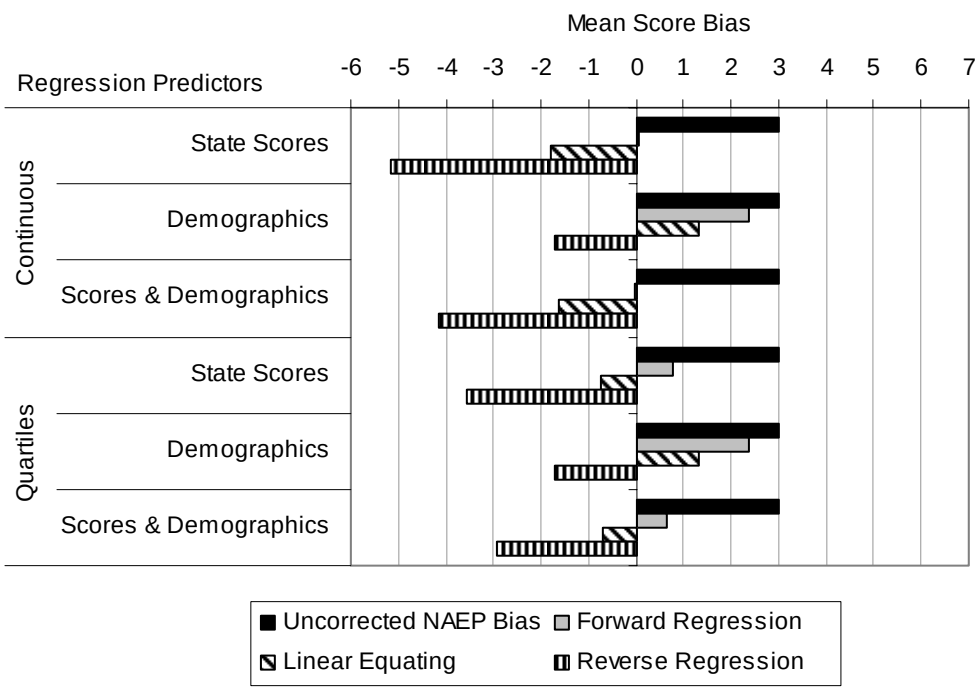
Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	-2.8	(1.2)	-0.2	(1.0)	0.8	(1.2)	2.3	(1.7)
	Demographics	-2.8	(1.2)	-1.3	(1.0)	-0.7	(1.2)	0.3	(1.6)
	Scores & Demographics	-2.8	(1.2)	-0.4	(1.1)	0.3	(1.2)	1.2	(1.6)
Quartiles of Predictors	State Scores	-2.8	(1.2)	-1.2	(1.0)	-0.2	(1.0)	1.7	(2.8)
	Demographics	-2.8	(1.2)	-1.3	(1.0)	-0.5	(1.2)	0.9	(1.8)
	Scores & Demographics	-2.8	(1.2)	-0.8	(1.1)	-0.1	(1.2)	1.0	(1.7)

Figure A9 – Grade 8 bias and bias corrections for non-participation of the 2 students in each school with the lowest NAEP scores, by predictor set and imputation method



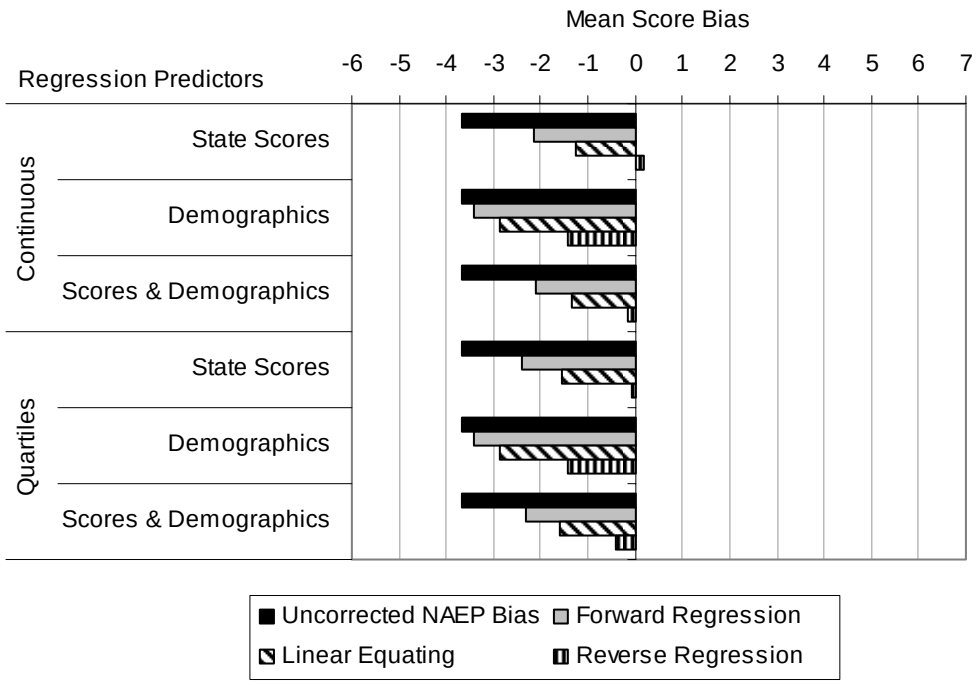
Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	4.8	(0.8)	3.1	(0.7)	2.2	(0.5)	0.5	(0.4)
	Demographics	4.8	(0.8)	4.1	(0.7)	2.7	(0.6)	-1.8	(1.1)
	Scores & Demographics	4.8	(0.8)	2.9	(0.7)	1.9	(0.5)	0.3	(0.5)
Quartiles of Predictors	State Scores	4.8	(0.8)	3.3	(0.7)	2.4	(0.5)	0.7	(0.5)
	Demographics	4.8	(0.8)	4.1	(0.7)	2.7	(0.6)	-1.8	(1.1)
	Scores & Demographics	4.8	(0.8)	3.0	(0.7)	2.1	(0.5)	0.4	(0.6)

Figure A10 – Grade 8 bias and bias corrections for non-participation of the 2 students in each school with the lowest state assessment scores, by predictor set and imputation method



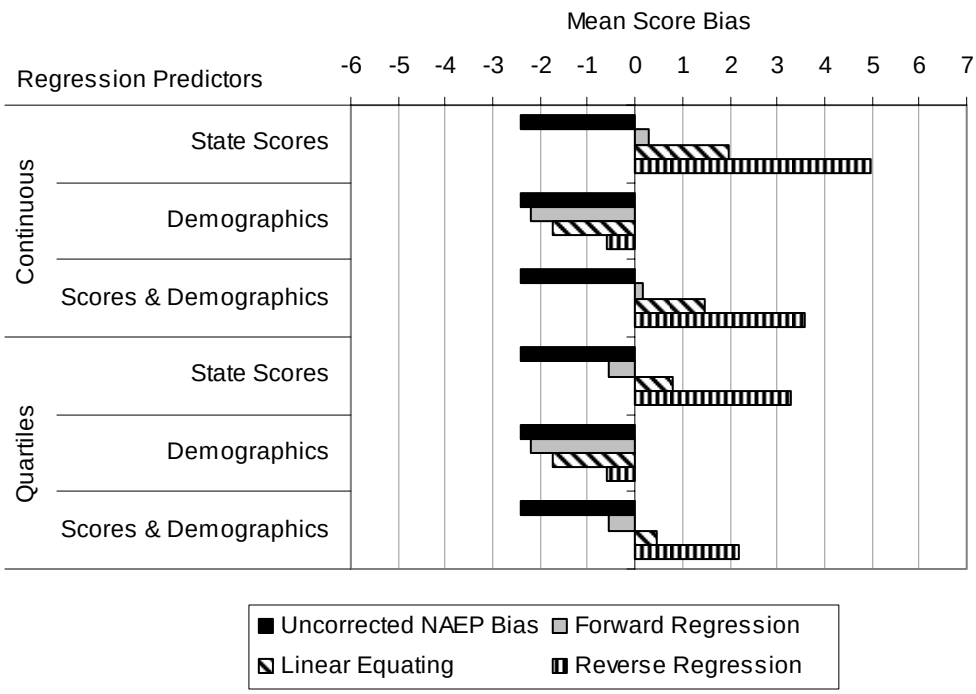
Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	3.0	(0.8)	0.0	(0.3)	-1.8	(0.8)	1.7	(2.9)
	Demographics	3.0	(0.8)	2.4	(0.6)	1.3	(0.5)	-2.1	(0.5)
	Scores & Demographics	3.0	(0.8)	0.0	(0.3)	-1.6	(0.6)	-4.2	(1.9)
Quartiles of Predictors	State Scores	3.0	(0.8)	0.8	(0.3)	-0.7	(0.4)	-3.6	(2.1)
	Demographics	3.0	(0.8)	2.4	(0.6)	1.3	(0.5)	-1.7	(0.5)
	Scores & Demographics	3.0	(0.8)	0.6	(0.3)	-0.7	(0.3)	-2.9	(1.3)

Figure A11 – Grade 4 bias and bias corrections for non-participation of the 2 students in each school with the highest NAEP scores, by predictor set and imputation method



Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	-3.6	(0.8)	-2.1	(0.7)	-1.3	(0.6)	0.2	(0.3)
	Demographics	-3.6	(0.8)	-3.4	(0.7)	-2.9	(0.6)	-1.4	(0.5)
	Scores & Demographics	-3.6	(0.8)	-2.1	(0.7)	-1.3	(0.6)	-0.2	(0.4)
Quartiles of Predictors	State Scores	-3.6	(0.8)	-2.4	(0.6)	-1.6	(0.4)	-0.1	(0.1)
	Demographics	-3.6	(0.8)	-3.4	(0.7)	-2.9	(0.6)	-1.4	(0.5)
	Scores & Demographics	-3.6	(0.8)	-2.3	(0.6)	-1.6	(0.5)	-0.4	(0.2)

Figure A12 – Grade 8 bias and bias corrections for non-participation of the 2 students in each school with the highest state assessment scores, by predictor set and imputation method



Mean Bias (Standard Deviation)		Uncorrected NAEP Bias		Forward Regression		Linear Equating		Reverse Regression	
Continuous Predictors	State Scores	-2.4	(0.7)	0.3	(0.3)	2.0	(0.9)	5.0	(2.7)
	Demographics	-2.4	(0.7)	-2.2	(0.7)	-1.8	(0.6)	-0.6	(0.7)
	Scores & Demographics	-2.4	(0.7)	0.2	(0.3)	1.5	(0.6)	3.6	(1.5)
Quartiles of Predictors	State Scores	-2.4	(0.7)	-0.5	(0.2)	0.8	(0.6)	3.3	(2.0)
	Demographics	-2.4	(0.7)	-2.2	(0.7)	-1.8	(0.6)	-0.6	(0.7)
	Scores & Demographics	-2.4	(0.7)	-0.6	(0.2)	0.5	(0.5)	2.2	(1.2)