Incorporating a First-Stage Finite Population Correction (FPC) in Variance Estimation for a Two-Stage Design in the National Assessment of Educational Progress (NAEP)

Jennifer Kali, John Burke, Lloyd Hicks, Lou Rizzo, Keith Rust
Westat, 1600 Research Boulevard, Rockville, MD 20850

Abstract

NAEP utilizes a two-stage design to select samples of school students in each state. Schools are selected with varying probabilities, via a stratified systematic sample. Students are selected with equal probability within schools. Jackknife replication is used for variance estimation, and the current procedure assumes that schools were selected with replacement. In many states the sampling fraction of schools is large, so that this approach overestimates the sampling error. We evaluated an approximate method for incorporating conservative finite population correction (FPC) factors directly into the replicate weights. Our investigation of the method involved two components: 1) an assessment of the impact of the change, using historic data; 2) an evaluation of the biases and variances of the past and proposed approaches, using a population generated to simulate the NAEP grade 8 reading population. We present the results of this evaluation.

Key Words: jackknife; school sample; replicate weights.

1. Introduction

Variance estimation procedures for the 2011 National Assessment of Educational Progress (NAEP) were changed from previous years regarding how the finite population correction (FPC) was incorporated. The 2011 JSM proceedings paper "Finite Population Correction (FPC) for NAEP Variance Estimation" by Rizzo and Rust describes the theory behind the changes. This paper is an evaluation of the new variance estimation method.

2. NAEP Overview

Conducted by the National Center for Education Statistics, the National Assessment of Educational Progress (NAEP) is a periodic assessment of student academic achievement which produces estimates at both the national and state levels. Assessments that include reporting at the state level are conducted bi-annually. These assessments are conducted on samples of fourth and eighth grade students, assessing reading, mathematics, and often science or writing.

The sample for these state-level assessments is a two-stage design, in which students are sampled within sampled schools. Samples are selected to be representative of the nation overall and for states and a select group of urban districts (denoted as the Trial Urban District Assessment (TUDA)). The reporting level (state or district) is commonly referred to as a jurisdiction.

Variances are estimated for NAEP using replication methods, specifically the jackknife with 62 replicates. Prior to 2011 (referred to from here on as the 'old method'), the variance estimation method did not include an FPC at the school level, but it did include an FPC at the student level for noncertainty schools only. The advantage of this approach was that was easy to implement in that it only required creating one set of replicates at the first stage sampling unit level (either at the school level for noncertainty schools or the student level for certainty schools).

Beginning in 2011 (referred to from here on as the 'new method'), the variance estimation method includes an FPC at the school level but not at the student level. As described by Rizzo and Rust in their paper, this corresponds to the agreed upon philosophy regarding FPC's for education studies, namely that schools are fixed elements that do not change and students are random elements that belong to a super population. There is some controversy surrounding this philosophy as it can be argued that students should have an FPC applied as well or, conversely, that there should not be FPCs at either level.

Note that the proportion of sampled schools within a jurisdiction varies greatly. In some jurisdictions all schools are included with certainty, in which case the FPC will not have an effect. The remaining jurisdictions vary from having as high as 83 percent of schools within a jurisdiction sampled to as low as nine percent of schools sampled. Jurisdictions with a large percentage of schools sampled will generally be more affected by including the FPC at the school level than jurisdictions with only a small percentage of schools sampled.

3. Evaluation

Before making the change to the variance estimation procedure, an evaluation was necessary to answer a few key questions. The first question was whether the new method makes appreciable difference to the NAEP results, compared with the old method. To answer this question, we reviewed the results of the 2009 NAEP study. The second question is whether the new method improves the variance estimation over the old method, meaning that it better represents the true variance while still remaining conservative. A simulation was required to answer this question.

3.1 Evaluation Using 2009 Data

Data from the 2009 NAEP study was used to address the first question regarding whether the new method makes an appreciable difference to the NAEP results. Standard errors were recomputed for both fourth- and eighth-grade reading and mathematics assessments. The standard errors from the new method were compared to standard errors from the old method in two ways. The first comparison was to review the variances of the weighted mean scores for students overall and within demographic subgroups (urbanicity, school size, race, sex). The second comparison was to reanalyze the published results to review differences between the new and old methods in terms of the statistical significance of various comparisons: trends from 2007 to 2009, subgroup comparisons within jurisdictions, and comparisons between jurisdictions and the nation overall.

This paper will present results from eighth grade only. There are fewer eighth grade schools in the population than there are fourth grade schools so the effect of the FPC on the standard errors for eighth grade was greater. The results for the fourth grade were consistent with those for the eighth grade.

There were 53 jurisdictions in the 2009 NAEP grade 8 assessment where not all schools were included in the sample. Table 1 shows the four jurisdictions with the largest sampling rates and the four jurisdictions with the smallest sampling rates. The proportion of schools included in the sample is indicated in the second column. Note that the jurisdictions with the largest sampling rates tend to be urban districts and the jurisdictions with the smallest sampling rates tend to be large states.

The table shows the standard errors of the weighted mean scores in reading of students overall, and within subgroups (black and female). These estimates are given as examples, though urbanicity, school size, race, and sex were all reviewed.

Table 1: 2009 Results – Standard Errors of Weighted Mean Reading Scores for Jurisdictions with Largest and Smallest Sampling Rates for the Old and New Variance Estimation Methods

	Pct schools	*Overall*		*Black students*		*Female students*	
Jurisdiction	*sampled within jurisdiction*	*SE (old)*	*SE (new)*	*SE (old)*	*SE (new)*	*SE (old)*	*SE (new)*
District A	82.76%	1.06	0.81	1.35	1.53	1.21	0.97
District B	78.72%	2.19	1.39	4.63	4.08	3.04	1.77
District C	75.41%	1.11	1.06	1.10	1.05	1.63	1.39
District D	71.91%	1.70	1.15	2.01	1.31	1.82	1.60
State A	13.55%	1.16	1.11	1.35	1.48	1.19	1.22
State B	12.85%	1.10	1.14	1.98	1.88	1.30	1.23
State C	9.33%	1.20	1.22	2.58	2.84	1.39	1.39
State D	8.83%	0.97	0.95	1.78	1.82	1.12	1.14

When the sampling rate is high, in general, the new estimate has a much smaller standard error than the old estimate. In some cases, the difference is quite substantial. For example, in District B, the standard error for the overall estimate decreases from 2.19 from the old method to 1.39 from the new method. This is not always the true, however, as in the case of the estimate for black students in District A, in which the standard error for the new method is larger than for the old method.

When the sampling rate is low, there is not an obvious pattern. Sometimes the old method has a smaller standard error while sometimes the new method has smaller standard errors. The effect of adding the FPC at the school does not have as strong of an effect when the sampling rate of schools within a jurisdiction is small.

Table 2 shows how the new variance estimation method affects the published significant results from the 2009 study. There are three types of published significant results. There are subgroup comparisons of the trend from 2007 and 2009 (e.g., black students in 2007 compared with black students in 2009), subgroup comparisons within 2009 (e.g., black students compared with white students in 2009), and comparisons of the nation to individual jurisdictions (e.g., State A compared with the US in 2009) or between jurisdictions (State A compared with State B). The rows of the table indicate how the new method of variance estimation changed the results. The first row indicates how many comparisons were not significant using the old method but became significant using the new method. The second row indicates how many comparisons were significant using the old method but were no longer significant with the new method.

Table 2: 2009 Results – Differences in Published Significant Results for the Old New Variance Estimation Methods

	Trend 2007-2009		*Subgroup comparisons within 2009*		*National public Vs jurisdictions within 2009*	
	Count	*%*	*Count*	*%*	*Count*	*%*
Total number of comparisons	2,454	-	3,599	-	2,080	-
Newly significant	15	0.61%	34	0.94%	27	1.30%
No longer significant	7	0.29%	3	0.08%	6	0.29%

In general, there are more newly significant results than no longer significant, which is what we would expect with smaller standard errors. However, the effect is small with the largest change being 1.30 percent.

The standard errors from the 2007 study were not recomputed using the new method. Therefore, the number of newly significant differences for trend comparisons is likely less than if standard errors for both 2007 and 2009 had been computed using the new method.

3.2 Evaluation Using Simulated Data

A simulation was required to answer the empirical question regarding whether the new method improves the variance estimation over the old method. The goal of the simulation was to examine the shape of the distribution of variance estimates to see if the new method was more stable and a better representation of the true variance while still remaining conservative.

A simulated NAEP population was derived by cumulating NAEP grade 8 reading data from 2002 to 2009 and propagating data to simulate the school population in every jurisdiction. 1,000 samples were drawn from each jurisdiction based on the sample design of the 2009 NAEP study.

Variances were estimated via new and old methods for each sample. The sampling variance of the simulation was computed (called ‘simulated true variance’ from here on), and compared the new and old estimated variances to it. The model utilized in creating the simulated data follows the variance estimation philosophy that schools are fixed entities.

The results of the simulation were evaluated based on bias, mean square error (MSE), and the shape of the distribution. The ideal distribution of a variance estimator would be chi-square with 62 degrees of freedom. The results of the simulations were reviewed for the estimates of students overall and within various subgroups (race, sex, English language learner status, student disability status, and eligibility for free or reduced price lunch).

Each of the 53 jurisdictions was categorized based on characteristics of the bias and mean-square error of the variance estimates derived from both the new and old methods. Table 3 summarizes the results for the overall estimate. This categorization was done for all of the subgroups, but only the overall results are presented here.

Table 3: Simulation Results – Summary of Bias and MSE of Variance Estimates for Overall Means

Characteristic of bias	*Smaller MSE*	*# of jurisdictions*	*% of jurisdictions*	*Mean bias: old method*	*Mean bias: new method*
Both overestimate variance, new method less biased	New	41	77.36%	0.67	0.41
Old overestimates variance, new method underestimates variance	New	6	11.32%	0.2	-0.15
	Old	1	1.89%	0.05	-0.14
Both underestimate variance, old method less biased	Old	5	9.43%	-0.81	-0.94

A large majority of the jurisdictions fell within the first category, in which both estimates overestimate the variance but the variance from the new method is less biased and has a smaller MSE. This is the best case scenario in that the new method remains conservative but is closer to the truth and has a tighter distribution.

Not all jurisdictions fell into this category, however. Twelve jurisdictions fell into categories that were more favorable for the old method of variance estimation. Seven cases fell into a category in which the old overestimates the variance but the new underestimates it. Since the preference is for a conservative estimator, the old method would be preferred in these cases. Five jurisdictions fall into the category in which both methods underestimate the variance but the old method is less biased.

Chart 1 is a plot of the root MSE of the old and new variance estimation methods for the overall estimate. The information presented is similar to what is shown in Table 3. The x-axis is the root MSE of the old method and the y-axis is the root MSE of the new method. Each plot point corresponds to a jurisdiction. The color of the plot indicates which method has a better bias, meaning either that it is positive whereas the other is negative, or else that it has smaller absolute value in the case when both have the same sign.

For the overall estimate, most jurisdictions fall below the 45 degree line. This indicates that, in general, the new method has a smaller MSE than the old method. Also, most jurisdictions are shown in blue, indicating that, in general, for the overall estimate, the new method has a better bias.

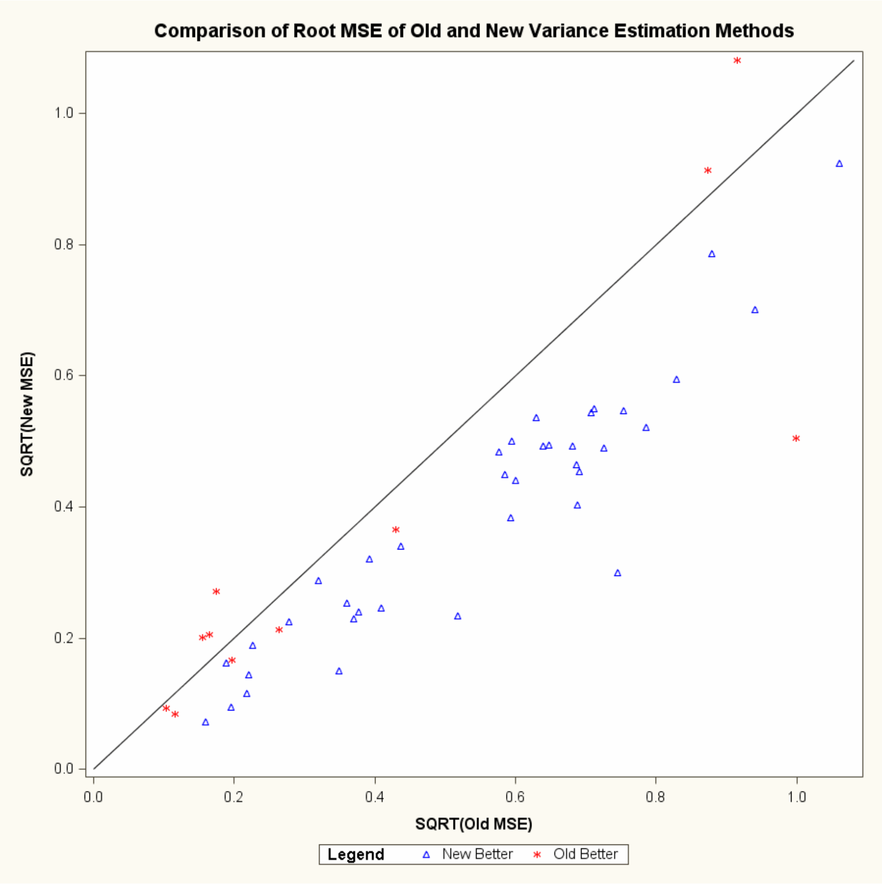

Chart 1: Scatter plot of root MSE of old and new variance estimation methods for overall estimate

The estimates shown in Chart 2 refer to students with disabilities (SD). This is a rather small subgroup of the student population. Note that the pattern of the scatter plot is different from that in Chart 1, which is for students overall. The jurisdictions tend to fall near the 45 degree line, and the numbers of blue and red jurisdictions are about equal. For small subgroups such as SD, including the FPC at the school level seems to have little effect overall. It seems likely that this is because students in small subgroups are less clustered within schools. Another factor may be that the old method tends to underestimate the student level component of variance, since the method is based on the assumption that a student level FPC applies, whereas in this application that is not the case. Thus the old method may in some cases have compensating biases, with overestimation of the school-level variance component and underestimation of the student-level component. This may contribute the number of cases where the old method has smaller overall bias than the new method.

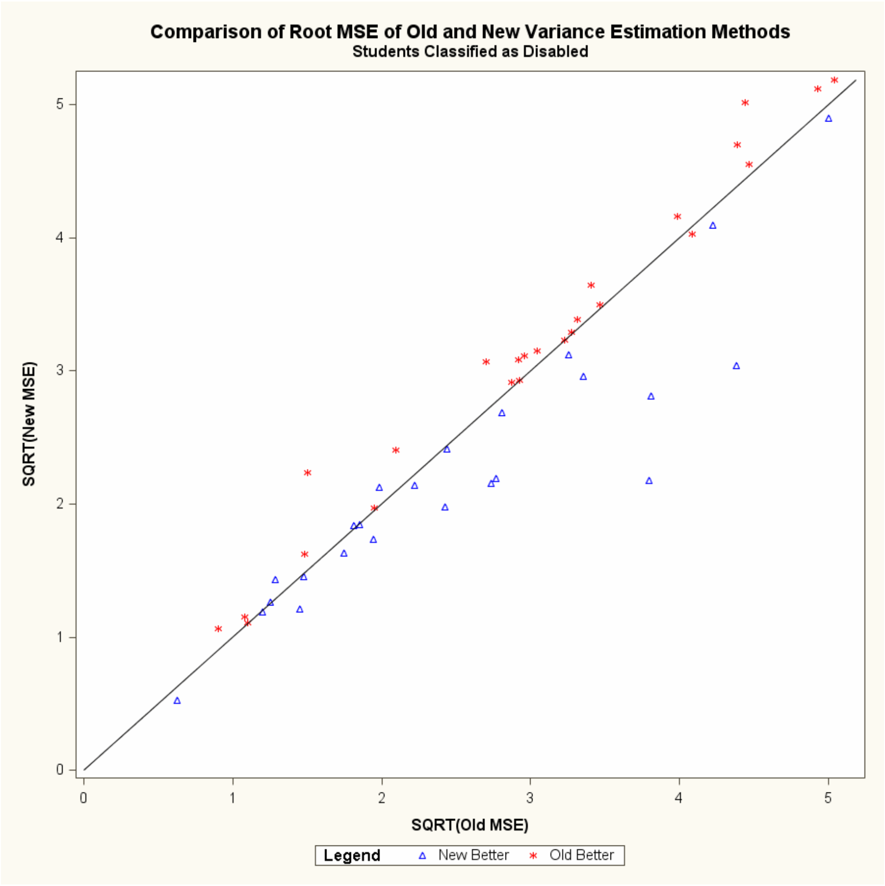

Chart 2: Scatter plot of root MSE of old and new variance estimation methods for estimate of students with disabilities

Histograms of the old and new variances estimation methods from each simulation run were reviewed for estimates of students overall and for the various subgroups (race, sex, English language learner status, student disability status, and eligibility for free or reduced price lunch) for each jurisdiction. Chart 3 is presented as an example. The histogram on the bottom shows the simulated variance estimates from the old method and the histogram on the top shows the simulated variance estimates from the new method. The vertical line indicates the simulated true variance. In the upper right corner of each plot are the values of the simulated true variance and the simulated mean of each variance estimation method.

The estimate for white students shown in Chart 3 is an example of a common pattern that emerged from the histograms. The new variance estimation method produces a much smoother distribution. Oftentimes the old method produces a distribution that is multi-modal while the new method produces a distribution that is much more similar to a chi-square distribution.

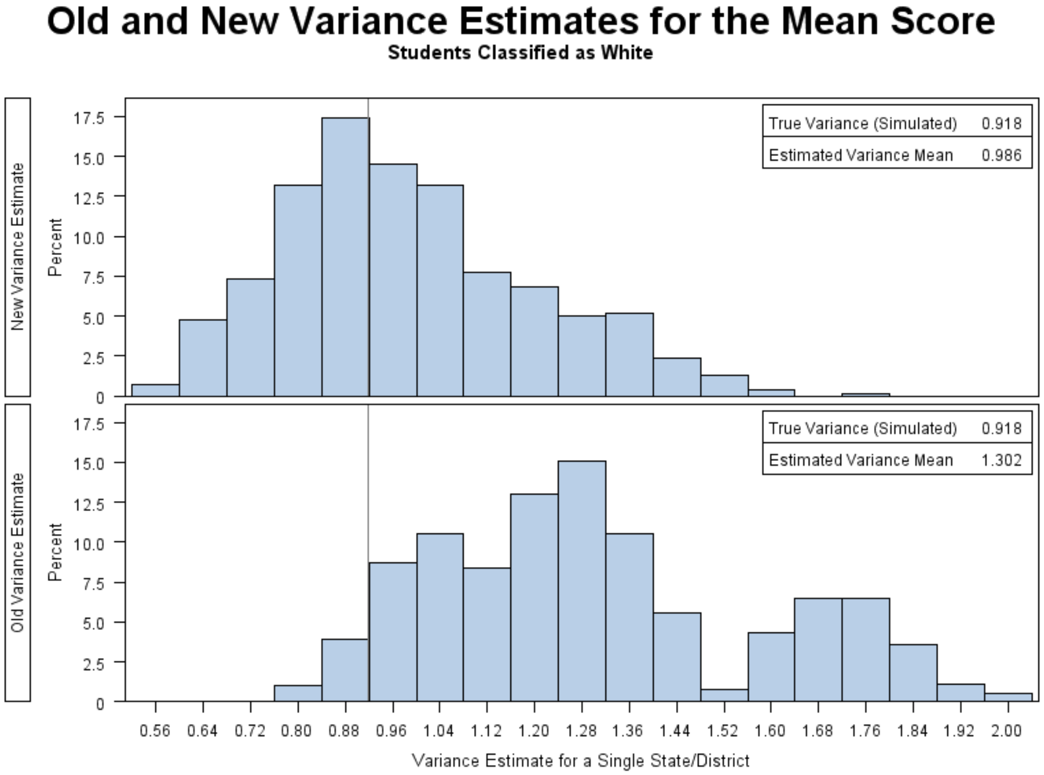

Chart 3: Histograms of simulated variance estimates of the mean score for the old and new methods for white students in a single jurisdiction

In general, for the majority of cases, the new method is less biased, has a lower MSE, and remains conservative. The distribution of variance estimates appears smoother and more closely resembles a chi-square distribution. The new method substantially lowers variances for overall means in a number of jurisdictions, especially urban districts.

However, the new method makes little difference for rare subgroups as students in these groups tend not to be clustered within schools. For rare subgroups, the new method often has slightly higher variance estimates than old method, presumably because the old method implicitly incorporated a student-level FPC (which we have deemed to be inappropriate).

4. Implementation

The results of the evaluation were presented to the NAEP Design and Analysis Committee in February 2011. A decision was made at that meeting that for the 2011 NAEP assessment variances would be computed using the new method. We also plan to repeat the evaluations done using the 2009 data, with 2011 data, comparing the new and old methods. Initial 2011 NAEP results are scheduled to be released in October 2011.

5. References

Rizzo, L. and Rust, K. (2011). Finite Population Correction for NAEP Variance Estimation. In *JSM Proceedings*, Survey Research Methods Section, Alexandria, VA: American Statistical Association.