Weight Calibration across Subject-Specific Samples in the National Assessment of Educational Progress

Jennifer Kali[1], Tom Krenzke[1], Keith Rust[1], John Burke[1]
[1]Westat, 1600 Research Blvd., Rockville MD, 20850

Abstract
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 level. Subsamples of the selected students are assigned to subjects such as mathematics and reading, and weighted totals of each subsample estimate the size of the student population. Previous standard weighting procedures for NAEP resulted in minor discrepancies in distributions of weight totals across demographic subgroups, between subjects. Raking (Iterative Proportional Fitting) was implemented in 2009 to eliminate the discrepancies in the demographic distributions. Subject specific weights were raked to sample-based population estimates from the entire sample. An evaluation was conducted to investigate the impact on the assessment means and standard errors.

Key words: Survey estimation, consistency of estimates

1. Introduction

Conducted by the National Center for Education Statistics (NCES), the National Assessment of Educational Progress (NAEP) is a periodic assessment of student academic achievement which produces estimates at both the national and state level. Subsamples of the selected students are assigned to subjects such as mathematics and reading, and weighted totals of each subsample estimate the size of the student population. Previous standard weighting procedures for NAEP resulted in minor discrepancies in distributions of weight totals across demographic subgroups, between subjects. Raking (Iterative Proportional Fitting), introduced by Deming and Stephan (1940) and discussed further in Oh and Scheuren (1987), was implemented in 2009 to eliminate the discrepancies in the demographic distributions. Subject specific weights were raked to sample-based population estimates from the entire sample. An evaluation was conducted to investigate the impact on the assessment means and standard errors. The details of the raking procedure and the evaluation are described in this paper.

2. NAEP Overview

The NAEP assessment of student performance, also known as 'The Nation's Report Card', is conducted bi-annually by the NCES. The assessments are conducted on a sample of fourth, eighth, and twelfth grade students on various subjects, including reading, mathematics, and science.

The sample 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 (called Trial Urban District Assessments (TUDAs)). Subsamples of students sampled within schools are selected to be assessed in particular subjects. In 2009 there were many concurrent assessments. We will be focusing in the paper on the fourth and eighth grade public school samples for the reading, mathematics, and science assessments. Science was optional for each state, though most participated. Estimates were produced at the national level, for each state, Washington, DC, and for 17 TUDA districts.

Weighted totals of each subject specific sample are estimates of the total student population. Therefore, the estimates of each subject within each state are each estimates of the same population, i.e. weighted totals of students in the sample for mathematics in fourth grade is a estimate of the total number of fourth grade students in Florida, as is the weighted total of the students in the Florida fourth grade reading sample. Prior to 2009, standard weighting procedures for NAEP resulted in minor discrepancies in distributions of weight totals across demographic subgroups, between subjects, due to random variations in student sampling.

3. Demographic Data Template

The Demographic Data Template is a report provided to each state and TUDA district showing weighted distributions of key demographic characteristics from the NAEP assessment within the state/TUDA district. The purpose of the report is for states/districtss to compare the distributions with data from the state or district to confirm the validity of the NAEP sample. Distributions are shown for gender, race/ethnicity, eligibility for free or reduced price lunch (an indicator of socio-economic status), student disability status, and English language learner status. Such reports are produced separately for each subject.

Minor variations in distributions between subjects can be seen if presented side-by-side. An example of the demographic data for fourth grade students in one state in 2009 is shown in Table 1 below. Results are shown for each subject mathematics, reading, and science.

Table 1: Demographic Data Template for 4th Grade Mathematics, Reading, and Science Students in One State in 2009

	Mathematics		*Reading*		*Science*	
Percent	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
GENDER						
Male	51.94	0.97	53.48	0.92	49.90	0.92
Female	48.06	0.97	46.52	0.92	50.10	0.92
RACE/ETHNICITY						
White	60.64	1.97	60.38	1.88	61.18	1.96
Black	33.10	2.02	32.91	1.89	33.17	1.98
Hispanic	4.37	0.60	4.38	0.64	3.91	0.64
Asian	1.10	0.29	1.36	0.41	1.23	0.25
American Indian	0.40	0.16	0.37	0.15	0.36	0.16
Other	0.38	0.09	0.59	0.16	0.15	0.05
SCHOOL LUNCH						
Eligible	54.63	2.05	54.78	1.94	54.29	2.05
Not eligible	45.37	2.05	45.22	1.94	45.71	2.05
SD/ELL						
Students with a disability	9.38	0.59	9.97	0.58	10.63	0.60
English language learners	2.44	0.42	2.22	0.33	2.36	0.43

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2009.

Notice the largest difference in this table is in the first row, for which the difference between subjects for gender is 3.6 percent between reading and science. Most differences are less than one percent. Since each subject is weighted to the same population, such discrepancies were potentially confusing to users of the data. Questions arose as to why there were more males sampled for reading than for mathematics and how that might bias the sample. Differences are simply due to sampling variations that arise from assigning subject types to students.

4. Details of the Raking Procedure as Applied to NAEP

Raking was proposed as the benchmarking solution for the cosmetic issue described above. As well as cosmetic agreement, we hoped to be able to reduce standard errors. Raking adjusts weights to match a population distribution known or based on a large sample. Student totals of acceptable quality, from external sources, do not exist to meet this need. It was proposed to rake the subject specific student samples to control totals computed from the nonresponse adjusted student sample prior to dividing the sample by subjects. The subject-specific student samples are each about a third of the larger student sample. The control totals are based on a sample that has a non-ignorable amount of sampling variance. With the replication-based variance approach used for NAEP, we were able to retain the magnitude of the sampling variance attributable to the larger sample by computing the control totals for the full sample and each replicate and by raking the full sample and replicate weights for each subject sample.

To explain raking, we first discuss poststratification. Poststratification involves one dimension of population subgroups; for example, gender is one dimension with two subgroups (male, female). The weights for males are adjusted such that the sum of weights equals the control total for males (and the same for females). A dimension can be formed by combining two variables. Since it was desirable to use several variables in the adjustment as it was applied to NAEP, the sample sizes associated with the resulting subgroup categories would be too small for a stable adjustment. The solution was to create several dimensions, one for each variable, and apply the poststratification procedure iteratively. The process began by first postratifying using the first dimension, then using the first iteration's adjusted weights, poststratify to the second dimension, and continuing until the maximum difference (between the sum of adjusted weights and the control totals) for each subgroup for each dimension was less than some pre-determined value.

The student weight is composed of the following factors: the school base weight, the school-level nonresponse adjustment, the within school student weight, the student-level nonresponse adjustment, and the subject adjustment, as shown in Figure 1. Excessively large subject adjusted weights are trimmed. Variance estimates are computed using the paired jackknife method with 62 replicate units. The raking process was implemented as the final weighting step. The control totals were summarized from the student sample files after the student nonresponse adjustment. The raking factors were applied to the weights with subject specific adjustment factor.

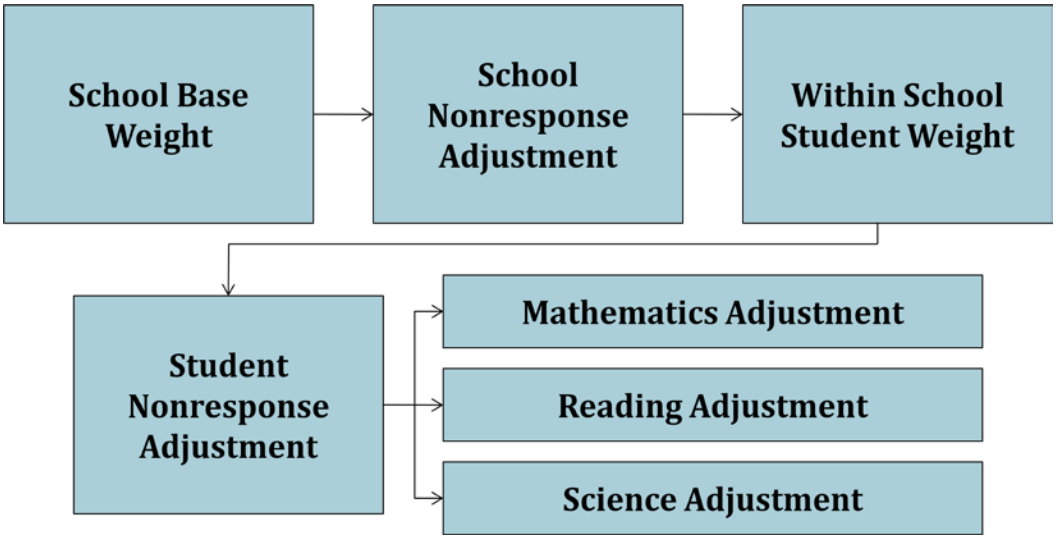

Figure 1: Weighting Steps for NAEP

Raking was conducted separately for each grade and jurisdiction. TUDA district and non-district balance of the jurisdiction were processed separately. There were 142 sets of control totals used in the raking. The process for each sample followed these steps:

1. Create the dimension variables on both the overall and subject-specific student data files (sex, race/ethnicity, Student disability status/English language learner status, school lunch eligibility).
2. Collapse raking dimensions according to rules regarding cell size.

3. Perform the raking procedure on the full sample and each replicate separately (62 replicates).

The raking dimensions were the same variables shown in the Demographic Data Template. Student disability status and English language learner status were combined into one dimension for raking though they are two separate variables.

Constraints were in place to prevent unstable raking adjustments. Categories of each dimension were combined whenever there were fewer than 30 responding students (20 for any of the replicates) in a subgroup. For simplicity, no restrictions were placed on the size of the adjustment factor. Collapsing was done as needed and thus may vary between subjects and grades.

5. Raking Adjustment Control Totals

The control totals are estimates of the student population within each raking dimension and are formed from the set of all assessed and excluded students. The control totals are computed as the weighted sum of students within each raking dimension as follows:

$$TOTAL_{dc} = \sum_{R_{dc}} STU\_BWT_{dck} \times SCH\_NRAF_k \times STU\_NRAF_{dck} \times SCH\_TRIM_k \times STU\_TRIM_{dck} / SUBJFAC_{dck}$$

where

- R_{dc} is the set of all assessed students in subgroup *c* of dimension *d*,
- STU_BWT_{dck} is the student base weight for a given student *k* in subgroup *c* of dimension *d*,
- SCH_NRAF_k is the school-level nonresponse adjustment factor for the school associated with student *k*,
- STU_NRAF_{dck} is thestudent-level nonresponse adjustment factor for student *k* in subgroup *c* of dimension *d*,
- SCH_TRIM_k is the school-level weight trimming factor for the school associated with student *k*,
- STU_TRIM_{dck} is the student-level weight trimming adjustment factor for student *k* in subgroup *c* of dimension *d*, and
- $SUBJFAC_{dck}$ is the subject factor for student *k* in subgroup *c* of dimension *d*.

The student weight used in the calculation of the control totals above is the adjusted student base weight, without regard to subject, reflecting nonresponse and trimming adjustments at the both the school and student levels. Control totals were computed for the full sample and each replicate independently.

6. Raking Adjustment Factor Calculation

For assessed and excluded students in a given subject, the raking adjustment factor STU_RAKE_k was computed as follows:

Initialize:

$$STUSAWT_k^{adj(4)} = STU\_BWT_{dck} \times SCH\_TRIM_k \times SCH\_NRAF_k \times STU\_NRAF_k \times SUBJFAC_k$$

Then,

For dimension 1:

$$STUSAWT_k^{adj(1)} = \frac{TOTAL_{dc}}{\sum_{R_{dc}} STUSAWT_k^{adj(4)}} \times STUSAWT_k^{adj(4)}$$

For dimension 2:

$$STUSAWT_k^{adj(2)} = \frac{TOTAL_{dc}}{\sum_{R_{dc}} STUSAWT_k^{adj(1)}} \times STUSAWT_k^{adj(1)}$$

For dimension 3:

$$STUSAWT_k^{adj(3)} = \frac{TOTAL_{dc}}{\sum_{R_{dc}} STUSAWT_k^{adj(2)}} \times STUSAWT_k^{adj(2)}$$

For dimension 4:

$$STUSAWT_k^{adj(4)} = \frac{TOTAL_{dc}}{\sum_{R_{dc}} STUSAWT_k^{adj(3)}} \times STUSAWT_k^{adj(3)}$$

After completing the adjustment for all four dimensions, if the maximum difference between the sum of adjusted weights and the control totals (for both full sample and replicates) for each subgroup for each dimension was less than some pre-determined value, then the process stops.If the stopping rule was not met, then the process proceeded by cycling back to the first dimension and continuing from there until the stopping rule was met. For the NAEP procedure, the pre-determined value for both the full sample and replicates was equal to 1.0.

Once the process converged, the adjustment factor was set equal to:

$$STU\_RAKE_k = \frac{STUSAWT_k}{STU\_BWT_{dck} \times SCH\_NRAF_k \times STU\_NRAF_k \times SCH\_TRIM_k \times STU\_TRIM_k \times SUBJFAC_k}$$

The process was done independently for the full sample and each replicate.

7. Results of the 2009 Implementation

A summary of the raking factors for the full sample is shown in Table 2 below. Each row is a summary of about 72 observations. The mean raking factor is close to one for each grade and subject. The factors all lie within the range (.535, 1.699) as highlighted in the table.

Table 2: Raking Factors for 2009 NAEP Assessment Main Sample

			Raking Factors	
Grade	*Subject*	*Mean*	*Min*	*Max*
4	Reading	1.002	0.535	1.390
4	Mathematics	0.999	0.661	1.699
4	Science	0.999	0.688	1.627
8	Reading	1.000	0.681	1.579
8	Mathematics	1.000	0.698	1.510
8	Science	1.000	0.584	1.586

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2009.

Table 3 shows the data from the demographic data for the same state shown previously. As this report was the motivation for raking, it is important to review the effect of the procedure on the variables in this report.

Table 3: Demographic Data Template for 4th Grade Mathematics, Reading, and Science Students in One State in 2009 after Raking

	Mathematics		*Reading*		*Science*	
Percent	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
GENDER						
Male	51.83	0.53	51.83	0.53	51.83	0.53
Female	48.17	0.53	48.17	0.53	48.17	0.53
RACE/ETHNICITY						
White	60.72	1.85	60.72	1.85	60.72	1.85
Black	33.05	1.89	33.05	1.89	33.05	1.89
Hispanic	4.26	0.58	4.26	0.58	4.26	0.58
Asian	1.08	0.34	1.18	0.29	1.37	0.31
American Indian	0.47	0.19	0.30	0.13	0.42	0.18
Other	0.41	0.12	0.48	0.11	0.17	0.06
SCHOOL LUNCH						
Eligible	54.58	1.88	54.58	1.88	54.58	1.88
Not eligible	45.42	1.88	45.42	1.88	45.42	1.88
SD/ELL						
Students with a disability	9.99	0.40	9.99	0.40	9.99	0.40
English language learners	2.30	0.34	2.36	0.34	2.34	0.34

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2009.

The percentages in the majority of rows in Table 3 match exactly between subjects, highlighting the success of the raking procedure. The first row, gender, the source of the largest discrepancy for this state prior to raking, is now an exact match.

There are still some lingering discrepancies. There are difference between subjects in the estimate of the percent of students who are English language learners. This variable was

combined with student disability status for raking, meaning that some control is lost for this individual variable.

There are minor differences for the smaller race/ethnicity categories (Asian, American Indian, and Other). These categories often have very small cell sizes that required collapsing. Each time collapsing is done, a little bit of the control is lost for the individual cells. The largest difference for this state is for the percentage of fourth grade students that are some other race. The difference is 0.31 percent, which is quite small.

8. Evaluation

Before finalizing this procedure, it was necessary to evaluate the overall effect of the raking procedure. Results using raked student weights were compared to results using unraked student weights for specific characteristics within the 71 sample jurisdictions and three additional national estimates for mathematics at each grade. For simplicity, only mathematics was used in this evaluation. The evaluation tried to determine

- the effect on selected demographic characteristics,
- the effect on mean mathematics scores,
- the effect on standard errors of mean mathematics scores.

8.1 Effects on Selected Demographic Characteristics

The raked weights were first evaluated to determine the effect of the weighting procedure on key demographic subgroups. We compared the sum of weights within specific student characteristics – the raking dimensions, relative age (modal for the grade, younger, older), school size (Large, Medium, and Small), and urbanicity (City, Suburban, Town, Rural) within the jurisdictions for each subject and grade. Among the tests conducted on the demographic characteristics, the proportion of p-values less than 0.05 is small. There were 116 out of 2,698 (4.3%) tests with p-values less than 0.05, which is well within the reasonable number expected. Therefore, the raking procedure does not have much of an effect on the overall demographic characteristics of the sample.

8.2 Effects on Mean Mathematics Score

Additionally, we compared the NAEP score before and after raking to see if the average score changed significantly after raking. Among the tests conducted on the scores, the proportion of p-values less than 0.05 is small. There were 111 out of 2,698 (4.1%) tests with p-values less than 0.05, which is well within the reasonable number expected. Therefore, the raking procedure did not have much of an effect on the overall scores. One thing to note is that tests with p-values less than 0.05 are clustered within a couple states. In only two states, at grade 8 over half of the 19 tests had p-values less than 0.05. Note that the tests within a state are not independent and some clustering was expected.

Another way to evaluate the effect on the scores is to compare the change in the score after raking to the standard error of the score before raking. Table 4 shows the subgroups for which the raking procedure changed the mean score by more than one standard error. This only occurred twice, both within the same TUDA district at grade 8. The two subgroups highly overlap since the proportion of students is large schools in a large proportion of the overall subgroup for this particular jurisdiction. If the absolute value of the ratio is greater than one, the raking procedure changed the values of the weighted scores by more than one standard error.

Table 4: Changes in Mean Mathematics Score by More Than One Standard Error

Subgroups	*Before Raking Average Score*	*Before Raking SE*	*Raked Average Score*	*Raked SE*	*Ratio Diff/ SE*
Overall	258.5	1.09	260.0	1.17	1.38
School Size: Large	261.4	1.51	263.7	1.59	1.52

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2007.

8.3 Effects on Standard Errors of Mean Mathematics Scores

Additionally, we reviewed the effect on the standard errors of the scores. Recall that we hoped to reduce our standard errors slightly by implementing the benchmarking procedure, but at the same time wanted to preserve the variability of the larger sample that was used for raking. The scatterplot in Figure 2 shows the standard errors of the mean mathematics score for eighth grade. Each point is a jurisdiction. The x-axis is the mean standard error of the jurisdiction after raking. The y-axis is the mean standard error of the jurisdiction before raking. It appears that the standard errors were slightly reduced by the raking procedure as the majority of the points (79%) are slightly above the 45 degree line.

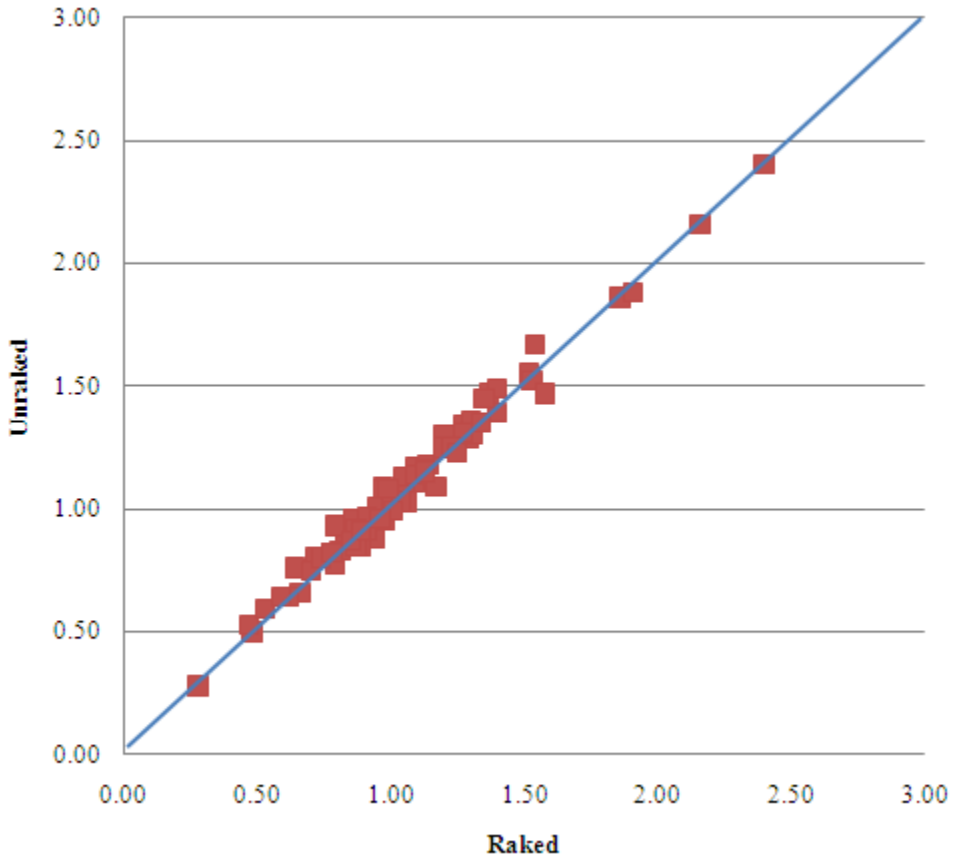

Figure 2: Scatterplot of mean standard errors for each jurisdiction before and after raking

9. Conclusions

The raking procedure implemented in the 2009 NAEP assessment weighting procedures appears to meet our goal of bringing consistency to the demographic distributions between subject estimates. Additionally, after careful evaluation, there do not appear to be any adverse effects on the mean assessment scores or standard errors of the mean assessment scores. Raking the full sample and replicate weights separately preserves the variability in the larger sample, therefore resulting in only a slight decrease to the standard errors for the subject samples due to adjusting to a larger sample. The raking procedure meets the needs of the assessment and will be used for future NAEP assessments.

References

Deming, W.E. and F.F. Stephan. 1940. On a Least Square Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known. Annals of Mathematical Statistics, 11: 427-444.

Oh, H.L. and F.J. Scheuren. 1987. Modified Raking Ratio Estimation. In *Survey Methodology*, 13: 209-219.