A COMPARISON OF RAKING AND POSTSTRATIFICATION USING 1994 NAEP DATA

Leslie Wallace and Keith Rust, Westat, Inc.
Leslie Wallace, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Keywords: Raking, poststratification, jackknife, forward regression

1. Introduction

The purpose of this paper is to compare the precision obtained using poststratification to that obtained using raking (also known as iterative proportional fitting), in a particular application. The 1994 National Assessment for Educational Progress (NAEP) will be used for this raking evaluation. NAEP is a congressionally mandated national assessment of students conducted every two years. In 1994, three distinct grade classes were targeted for the assessments: grades 4, 8, and 12. Students were assessed in reading, history, and geography at each grade class. The samples were selected using a complex multistage sample design involving the sampling of students from participating schools within 94 selected geographic areas, called primary sampling units (PSUs), across the United States. Private schools and schools containing moderate to high numbers of Black or Hispanic students were oversampled in order to yield enough students for analysis in these subgroups. In 1994, about 90,000 students were selected for the three grades and the three assessment subjects. The assessments were conducted in the spring of 1994. For more information about the NAEP assessment, see the *1994 NAEP Technical Report* (1996).

Historically, NAEP weights have been poststratified to control totals by age, grade, race, and region. These control totals are derived using data from the October education supplement to the Current Population Survey (CPS). While poststratification reduces the variances of NAEP survey estimates by as much as 50 percent over what would be achieved using non-poststratified weights, it was postulated that raking instead of poststratification might reduce the variances even further. This is because raking would control the distribution of the final sample weights with respect to a greater variety of variables, related to educational achievement, than can be achieved through poststratification. For a description of the technique of raking in survey estimation, see Oh and Scheuren (1987).

The evaluation focuses on the 1994 reading assessment because it is the largest, and presumably yields the most stable results. About 37,200 students were assessed in reading: 7,382 at grade 4; 15,606 at grade 8; and 14,181 at grade 12. Fourteen cells were used to poststratify students in the reading assessment in 1994. The cells were formed by crossing two eligibility classes of (modal age, not of modal age) with seven levels of race/region (White-Northeast, White-Southeast, White-Central, White-West, Hispanic, Black, Other race). The definition of modal age was based on age in years as of October 1, 1993. Students aged 9 were considered of modal age for grade 4, students aged 13 were considered of modal age for grade 8, and students aged 17 were considered of modal age for grade 12. Age was defined as of October 1, so as to be consistent with the age definition used in the annual October education supplement to the CPS.

The primary analysis statistic used in the NAEP assessment is a proficiency score for each student. For information about NAEP proficiency scores, see the *1994 NAEP Technical Report* (1996). The raking evaluation will compare estimated standard errors and coefficients of variation for mean reading scores within certain subgroups of interest using both poststratified and raked weights at each age class. The subgroups of interest are those traditionally published by the National Center for Education Statistics in standard NAEP reports such as *NAEP 1994 Reading: A First Look* (1995). They include region, race, gender, parents' education level, and school type.

The general approach to conducting this evaluation involved several steps. First, variables suitable as candidates for the raking process were chosen. Then how these variables would be combined to form raking dimensions was determined. Next, control totals were calculated. Raked weights were created for students assessed in reading at each grade. For each grade, both raked estimates and poststratified estimates were calculated and compared. Each of these steps is described below in more detail.

Finally, the raking and poststratification, and the comparison between them, were repeated for the history and geography assessments for each grade.

2. The Choice of Candidate Variables for Raking

The variables that might be used in raking were limited to items known or thought to be correlated with student performance on the assessments, to what information was known for respondents in the NAEP assessment, and by the level at which student counts were available to form control totals. With these restrictions in mind, the variables chosen as candidates for the evaluation include many of the reporting subgroups of interest, and many student-level and school-level characteristics known to be correlated with student performance on the assessments. In addition, these variables and estimates of numbers of students are available from the educational supplement to the CPS, so that control totals can be calculated.

The following variables were used in the evaluation: *age* (old for the grade, modal for the grade, young for the grade); *race* (White and other, Black,

Hispanic, Asian and Pacific Islander, American Indian); NAEP *region* (Southeast, Northeast, Central, West); *metro status* (MSA, non MSA); and *school type* (Public, Private). Students in grades 4, 8, and 12 who were assessed in reading in the 1994 NAEP assessment were assigned values for these variables. As discussed, *age* was based on an October age definition, since that is the definition used by CPS. For example, for grade 4, students born before 10/83 were considered "old", students born from 10/83 to 9/84 were considered "modal", and students born after 9/84 were considered "young". NAEP regions differ somewhat from Census regions but are (in large part) defined as aggregates of states (see the *1994 NAEP Technical Report* for a definition of NAEP regions).

Two reporting variables not included in the evaluation are *gender* and *parents' education level*. *Parents' education level* was not considered due to the inability to find a suitable source for control totals. *Gender* was not considered because for reading (and a number of other subjects assessed by NAEP), differences in achievement by gender are very small. Another variable of interest that could not be included in the evaluation was a student's eligibility for the free lunch program. This variable was discounted because it was being collected in NAEP for the first time in 1996, and thus could not be part of an evaluation using 1994 data. It may be a candidate for future evaluations.

3. The Definition of Raking Dimensions

The variables *age*, *race*, *region*, *metro status*, and *school type* were analyzed for the 1994 NAEP reading data using regression techniques to see which were good predictors of reading proficiency scores. Both main effects and two-way interaction terms were included in the models. Each significant main effect or interaction term potentially defined a raking dimension. A significant *age x race* interaction term, for example, would lead to forming one dimension for raking by crossclassifying *age* and *race*. The analysis was done separately at each grade. It was highly desirable to limit the number of dimensions to three or four, and to use the same dimensions, and thus the same raking procedures, at each grade. This simplicity would make implementing the raking procedures in future rounds of NAEP compatible with the relatively short schedule allowed for weighting the assessment data. In addition, simplicity would improve the robustness of the raking procedure by keeping it from being too dependent on the assessment chosen for the evaluation (1994 reading), while hopefully still providing gains in precision over poststratification.

In addition to the five main effects, six two-way interaction terms were identified and included in the evaluation: *age x race*, *age x region*, *age x school type*, *race x region*, *race x school type*, and *race x metro status*. The model is given by:

$$Y_i = \beta_0 X_0 + \beta_1 A + \beta_2 R + \beta_3 G + \beta_4 S + \beta_5 M + \beta_6 AR + \beta_7 AG + \beta_8 AS + \beta_9 RG + \beta_{10} RS + \beta_{11} RM + \varepsilon_i \quad (1)$$

where

Y_i = the reading proficiency score for student i;

A = the age main effect, (two indicator variables);

R = the race main effect, (for indicator variables);

G = the region main effect, (three indicator variables);

S = the school type main effect, (one indicator variable);

M = the metro status main effect, (one indicator variable);

AR, AG, AS, RG, RS, and RM, are two-way interactions, (36 indicator variables in total); and

ε_i = residual term.

Two different software packages were used for the evaluation: SAS and WesVarPC. SAS can handle larger models, and has more options such as forward regression procedures. However, the primary disadvantage of SAS is that standard errors are based on a simple random sampling assumption, which is not appropriate for complex surveys such as NAEP. The standard errors under simple random sampling are substantially smaller that those expected from the NAEP design. WesVarPC is estimation software that was developed at Westat, Inc. for complex surveys. WesVarPC includes limited regression analysis capabilities. Regression parameters are estimated using the method of weighted least squares. Variances are estimated using replication methods including jackknife, which is the variance estimation method used for NAEP. The basic approach for the raking evaluation was to use SAS to develop initial models, and then refine them in WesVarPC.

First a forward regression procedure was run in SAS using the model in Equation (1). Indicator variables were created for each level of each main effect and interaction term in the model. The main effects were forced into the model because we were interested in keeping interaction terms (i.e., crossing variables to form raking dimensions) only if they explained something that the main effects did not. Individual indicator variables that represented specific levels of interaction terms were added to the model at the .01 level. Seven indicator variables were added to the model at grade 4, seven were added at grade 8, and four were added at grade 12, as shown in Table 1. Thus, *age x region* is represented at grades 4 and 8, *race x region* and *race x school type* are represented at all three grades, *race x metro status* is represented at grades 4 and 8, and *age x race* is represented at grade 8 only.

Table 1. Significant interaction terms from the SAS forward stepwise regression procedure

Indicator variable	Interaction
Grade 4:	
old for the grade, NE	*AG*
old for the grade, SE	*AG*
Blacks, SE	*RG*
Asian and Pacific Islanders, NE	*RG*
Blacks, private schools	*RS*
Hispanics, private schools	*RS*
Hispanics, non MSAs	*RM*
Grade 8:	
old for the grade, SE	*AG*
young for the grade, SE	*AG*
Native American, old for the grade	*AR*
Hispanics, NE	*RG*
Blacks, private schools	*RS*
Hispanics, private schools	*RS*
Asian and Pac. Islanders, non MSAs	*RM*
Grade 12:	
Blacks, NE	*RG*
Blacks, SE	*RG*
Blacks, private schools	*RS*
Hispanics, non MSAs	*RM*

The models that resulted from the forward procedure were run in WesVarPC to see whether any terms would drop out once the improved estimates of standard error were used. Variables with p-values larger than .05 were dropped. One *race x region* term and one *race x school type* term dropped out at grade 4, but other terms for these interactions remained, so that the relevant set of interaction terms as a whole was not eliminated in each case. The *race x metro status* term at grade 8 dropped out, so this set of interaction terms was eliminated. No terms dropped out at grade 12. Thus the reduced model at grade 8 included four sets interaction terms instead of five.

In the interest of simplicity and robustness (discussed earlier) it was decided that all levels of an interaction term would be considered for raking (sample size permitting) if any one of the associated indicator variables was significant. Also, using the same raking dimensions at each grade was desirable. Therefore, a new model was run in WesVarPC that included the main effects and all indicator variables for the five sets of interaction terms that had been significant in the previous run for at least one grade. The new model run at each grade was the same as that in Equation (1), except that the *age x school type* interaction term *(AS)* was dropped. In addition, simultaneous hypothesis tests were done on each set of indicator variables for each interaction term to see if any interaction terms could be dropped from the model. The goal was to obtain interaction terms that could consistently be left out of the model for all three grades. The remaining interaction terms would form the basis for constructing raking dimensions. The results of these hypothesis tests are shown in Table 2.

Table 2. P-values for testing whether beta = 0 for all levels of a given interaction

	P-value		
Interaction	Grade 4	Grade 8	Grade 12
AG	.0685	.0222	.6867
RG	.1137	.0058	.0036
RS	.0581	.0001	.0000
RM	.1120	.4993	.9821
AR	.1065	.0085	.9752

A decision was made to drop interaction terms that failed to reject the null hypothesis at the .01 level for at least one grade. Thus, *age x region, race x metro status*, and *age x race* were not chosen as raking dimensions. The dimensions chosen were *race x region, race x school type, age*, and *metro status*. Thus all of the original main effects that were significant predictors of reading proficiency scores were accounted for in the raking.

The variables chosen as raking dimensions also explain some differences in response rates at both the school and student level. This was confirmed graphically, but formal statistical tests were not done for response rates. Many of the variables that will be used to form the raking dimensions are not used in forming nonresponse adjustment cells. Thus, the fact that the variables used in raking are associated with response rates introduces the possibility that the raking procedure will reduce the potential for bias due to nonresponse, in addition to improving precision. Note, however, that we have not attempted to evaluate any bias reduction due to raking, nor to compare it with poststratification for this aspect.

4. The Calculation of Control Totals

Control totals were calculated for each raking dimension at each grade using the October 1993 Education Supplement to the CPS. The totals needed for each grade were: race by region (20 cells), race by school type (20 cells), age (three cells), and metro status (two cells). The control totals were obtained by 1) extracting the necessary information from the CPS file (processing was restricted to CPS respondents in grades 4, 8, or 12.) 2) assigning CPS respondents to each raking cell, and 3) producing weighted estimates of the number of students in each raking cell. These estimates were the control totals.

5. The Calculation of Raking Factors and Raked Weights

For each cell of the four way table formed by the raking dimensions, the quantity $\hat{n}_{co}$ was calculated as the sum of the student weights of students who were

members of that cell. This was also repeated using the 62 sets of jackknife replicate weights to derive replicate cell totals $\hat{n}_{cj}\left(j=1,...,62\right)$. These cell totals were then raked to the independent marginals via the method of Iterative Proportional Fitting. Let the final raked cell totals be denoted as $N_{cj}\left(j=0,...,62\right)$. Then, for each cell, the raking factor, and its replicates, were calculated as $f_{cj}=N_{cj}/\hat{n}_{cj}$, $\left(j=0,...,62\right)$.

Raked weights were computed for each student assessed in reading in the relevant grade (4, 8, or 12) as follows:

$$W_{ij}=W_{ij}^{'}\times f_{cj}$$

where

W_{ij} = the raked full-sample (j=0) or replicate (j=1,...,62) weight for student i,

$W_{ij}^{'}$ = the NAEP student weight before poststratification, incorporating nonresponse and trimming adjustments,

f_{cj} = the full-sample (j=0) or replicate (j=1,...,62) raking adjustment factor for students in adjustment cell c.

The raking was done for each grade separately for the full-sample and each replicate across the four dimensions *(race x region, race x school type, age*, and *metro status*) until convergence was met.

6. The Production of Poststratified and Raked Estimates

Standard errors and coefficients of variation for mean reading proficiency scores, overall and within subgroups often used in NAEP reporting, were computed using WesVarPC. The jackknife method of variance estimation was used, using the set of 62 replicate weights described above. The subgroups used were region (Northeast, Southeast, Central, West), race (White, Black, Hispanic, Asian, Pacific Islander, American Indian), gender, parents' education level (graduated college, some education after high school, graduated high school, did not finish high school, don't know), and school type (Public, Catholic, Other Private). The estimates were computed separately using raked and poststratified weights at each grade.

Estimates were also computed using weights subject to neither raking nor poststratification. These were used as a baseline for measuring the relative effectiveness of the raking and poststratification procedures.

7. Results and Conclusions

In considering the results of the analysis, we first consider whether earlier findings, that the current NAEP poststratification procedure leads to significant gains in precision, are supported for the 1994 results in Reading, for all three grade levels. We then consider the effectiveness of the raking procedure we have developed for giving improvements in precision, both in absolute terms, and in comparison with poststratification. We then consider whether these results are replicated for the Geography and History assessments. Finally, we consider the implications of the results.

Table 3 shows the precision of estimates of mean reading proficiency for each grade for the whole population and a variety of demographic subgroups. The table shows the size of the estimated standard error for poststratified weights, relative to a baseline set of weights which are neither raked nor poststratified. It also shows the size of the estimated standard errors using raked weights relative to the baseline weights, and finally, the standard errors from raking relative to those from poststratification. In each case the number in the table shows, in percentage terms, the standard error from the first procedure, minus the standard error from the second procedure (with which the first is being compared), divided by the standard error from the second procedure. Thus a negative quantity indicates that the first procedure has given a smaller standard error estimate. The table also shows the results for geography at grade 8, which we discuss later.

Looking at the first column for reading for each grade, it can be seen that poststratification has given a significant reduction in sampling error for the overall mean at grades 4 and 12, but no noticeable gain at grade 8. When subgroups are considered, benefits can be seen for grades 4 and 12 for estimates by region, gender, metro status, and for public school students. For Parents' Education, there in general appear to be benefits of poststratification for these two grades, but the results vary somewhat by subgroup and grade for this classification.

For grade 8, however, there is no evidence of any consistent benefit from poststratification, and in fact in a few cases the poststratified estimates have substantially higher standard errors. This variation for grade 8 from the other two grades, and previous NAEP experience, is puzzling. We consider below whether this phenomenon holds in the results for History and Geography. The results for grades 4 and 12 are rather as we would have expected. As the dominant poststratification variable is race/ethnicity, it is not surprising that there are important gains through poststratification for categories that cut across race/ethnic groups, but these are much diminished for the various race/ethnic subgroups. It does not appear that the poststratification by age has given significant reductions at grade 4 for race/ethnic groups. The equivalence of poststratified and baseline weights for non-white race/ethnic groups at grade 12 is structural, since there are no age group poststrata at this grade, and region is not used for non-white groups. The use of region may give some modest benefit, as indicated by the 7 percent reduction in standard error for grade 12

white students, where the only poststratification variable in effect is region.

Turning to the raked weights, in undertaking this research we had expected that raking might perform somewhat better than poststratification for whole population estimates, since it uses more variables and an effort was made to optimize the way in which the variables were used. We also thought that it was possible that raking would give significant improvement for those subgroups where poststratification did not give any benefits.

An examination of the second and third columns for each grade show that these hopes were largely unrealized for Reading. At grades 4 and 12, raking gave good gains for the whole population, but not better than poststratification. At grade 8, where poststratification failed to give benefits, raking gave somewhat smaller standard errors. But again, as the results for poststratification were so unexpected for this grade, it is important to consider the results for History and Geography before attempting to interpret this finding.

At grade 4 and 12, raking generally gave similar or poorer results than did poststratification, for population subgroups. This was even true for race/ethnic groups, where poststratification gave little benefit, and had no effect for non-whites at grade 12. The raking appears to have given some benefit for results for Hispanics, Pacific Islanders, and Native Americans at grade 12, but done nothing for Black and Asian student results. At grade 4, raking has almost universally performed slightly worse than poststratification. Thus the results for Reading suggest that raking may be of some benefit at grade 12, where poststratification by age was not possible, and at grade 8, where poststratification was ineffective. We now consider whether these findings generalize to Geography and History, especially at grade 8.

The right-hand set of columns in Table 3 show the results for Geography for grade 8. The results for both Geography and History at grades 4 and 12 were quite similar to those for Reading, and at Grade 8 the results for History were quite similar to those for Geography. Comparing the results for Reading and Geography at grade 8, however, shows that very different results were achieved. The results for Geography are much more consistent with the findings at the other grades. Both raking and poststratification give a very significant reduction in standard error overall, but are very comparable to one another. The benefits are largely diminished for population subgroups, except those defined by region, and for public schools. Raking gives a noticeable improvement over poststratification only for regional estimates, a result which is not replicated with any consistency when considering the results for grades 4 and 12, and for History.

This leads to the question as to why raking was not able to lead to greater improvements than poststratification, when raking was able to incorporate control on more variables, each of which was found to be significantly related to achievement, at least in reading. We speculate that there are two reasons for this. The first is that the traditional NAEP poststrata remain good ones, capturing most of the variance that can be explained by the candidate variables for raking. In particular, the use of race/ethnic classes captures a very large portion of the total explainable variance in its own right. The second explanation is that those variables included in the raking, but not the poststratification, are actually well controlled by the sample design. Thus although school type and metro status appeared as significant variables in the models, and are absent from the poststratification variables, they are in fact well controlled by the stratification and systematic sampling procedures used in drawing NAEP school samples. Thus there is in fact little remaining variance associated with these variables to be removed by raking.

This in turn suggests that region may be of limited use as a variable for raking and poststratification, since it too is well controlled by stratification. It seems that the greatest benefits can accrue by using race/ethnicity and age; student characteristics rather than school characteristics, which therefore are less amenable to control at the design stage. Since the current poststratification procedure cross classifies race/ethnicity by age, it seems likely that it is capturing most of remaining variability that can be explained by the candidate variables.

In summary, although implementing a raking procedure on a routine basis as part of the NAEP weighting procedure would be a relatively straightforward task, this analysis has not produced any convincing evidence that such a change would give benefits that would justify the slight increase in effort that would be required.

REFERENCES

Oh, H.L., and Scheuren, F. (1987). Modified raking ratio estimation. *Survey Methodology*, 13, 209-219.

Schaffer, J., and Carlson, J.E., (eds). (1996). *The 1994 NAEP Technical Report.* National Center for Education Statistics: Washington, D.C.

Williams, P.L., Reese, C.M., Campbell, J.R., and Phillips, G.W. (1995). *NAEP 1994 reading: A first look. Findings from the National Assessment of Educational Progress.* National Center for Education Statistics: Washington, D.C.

Table 3. Relative standard errors of mean proficiency score by subject, grade and weighting adjustment method:

	Reading Grade 4			Reading Grade 8			Reading Grade 12			Geography Grade 8		
	Poststr. vs. baseline	Raking vs. baseline	Raking vs. poststr.	Poststr. vs. baseline	Raking vs. baseline	Raking vs. poststr.	Poststr. vs. baseline	Raking vs. baseline	Raking vs. poststr.	Poststr. vs. baseline	Raking vs. baseline	Raking vs. poststr.
Overall:	-27.2	-19.4	+10.6	-1.4	-9.8	-8.6	-26.2	-26.4	-0.3	-41.7	-41.2	+1.0
Race:												
White	-3.4	+4.8	+8.5	+8.3	+1.0	-10.0	-7.7	-8.0	-0.3	-7.2	-5.4	+1.9
Black	-2.9	+11.9	+15.2	+6.7	+1.7	-14.0	0.0	+1.1	+1.1	+1.8	-10.6	-12.2
Hispanic	-2.6	+17.4	+20.5	-7.1	+1.2	+4.6	0.0	-11.1	-11.1	-10.6	-6.4	+4.8
Asian	+1.6	+5.6	+3.9	+17.5	+3.0	-7.5	0.0	-0.9	-0.9	+3.1	-7.7	-10.5
Pac. Islander	-0.8	+16.5	+17.4	-2.0	+6.6	-7.1	0.0	-8.1	-8.1	+0.9	-15.6	-16.3
Native Amer.	0.0	+18.2	+18.2	+18.1	+3.9	-16.6	0.0	-29.1	-29.1	-5.3	-15.1	-10.3
Other	-0.8	-1.8	-1.1	+24.5	+11.4	-23.0	0.0	+4.6	+4.6	-0.8	+3.1	+4.0
Region:												
Northeast	-18.9	-20.0	-1.3	+14.0	+2.4	-16.7	-8.4	-5.3	+3.3	-9.6	-23.7	-15.6
Southeast	-23.5	-22.7	+1.0	-16.5	+1.8	-16.9	-5.7	-6.9	-1.3	-31.7	-42.0	-15.1
Central	-8.0	-5.7	+2.6	+24.3	+1.7	-5.6	-17.8	-35.4	-21.4	-16.2	-22.7	-7.8
West	-26.3	-31.7	-7.4	+7.9	+1.4	-16.8	-30.6	-37.8	-10.3	-36.5	-39.9	-5.3
Parents' ed.:												
Less than HS	+2.2	+9.8	+7.4	+4.0	+1.8	+2.8	-10.9	-6.8	+4.6	-9.6	+2.6	+13.4
Graduated HS	-13.5	-1.5	+13.8	+9.9	+1.2	-17.8	+2.9	-1.0	-3.8	-22.0	-23.9	-2.3
Post HS	-2.6	+0.4	+3.0	+28.4	+1.4	-18.1	-6.8	-0.8	+6.4	-13.6	-11.9	+2.0
Grad. college	-18.4	-12.0	+7.9	+19.5	+1.0	-12.6	-16.4	-17.6	-1.5	-17.7	-12.4	+6.4
Unknown	-22.0	-10.8	+14.3	+2.1	+1.5	-12.1	+6.8	-12.0	-17.6	-23.1	-22.4	+0.9
School type:												
Public	-26.3	-22.9	+4.6	-12.3	+0.9	-4.6	-28.9	-26.4	+3.5	-42.0	-42.4	-0.7
Catholic	+1.8	-22.8	-24.2	+34.9	+1.5	-25.4	+0.3	-0.4	-0.6	+7.9	+0.5	-6.8
Private	+9.5	+18.9	+8.6	+0.4	+2.2	-1.8	-4.1	+6.3	+10.8	-6.9	-6.0	+1.0
Sex:												
Male	-17.4	-12.3	+6.2	-0.5	+0.9	-5.3	-13.1	-19.6	-7.4	-34.4	-34.4	0.0
Female	-24.4	-16.2	+10.8	+9.3	+1.0	-19.0	-28.3	-21.3	+9.8	-36.4	-34.2	+3.5
Urbanicity:												
Large MSA	-6.9	-3.0	+4.2	+27.0	+1.4	-7.1	-1.9	-1.2	+0.7	-5.3	+3.1	+8.8
Other MSA	-13.3	-6.0	+8.5	+9.0	+1.7	-10.8	-17.5	-6.9	+12.8	-14.8	-13.6	+1.5
Non MSA	-11.8	-9.4	+2.8	-0.6	+1.9	+4.0	-12.6	-9.9	+3.1	-2.7	-7.5	-4.9