Analysis of Design Effects for NAEP Combined Samples

Jiahe Qian and Bruce Kaplan
Jiahe Qian, Educational Testing Service, MS 02-T, Rosedale Road, Princeton, NJ 08541

Key words: merging samples, complex sampling, weighting, relative precision

This proposed National Assessment of Educational Assessment (NAEP) study analyzes the efficacy of merging national and state NAEP samples. The goal of merging the two samples is to provide more accurate estimations, especially for small groups such as SD/LEP students. The study is based on combined samples of the 1998 NAEP 8th Grade Reading Assessments. The National reading reporting sample contains 11,193 students, and the State reading reporting sample contains 91,206 students, making the combined sample size 102,399. Both national and state assessments used the same instruments and were administered with similar procedures. Section 1 describes the combining procedure.

The study addresses statistical factors that influence design effects in educational assessments. Of the factors that can influence design effects, such as stratification, multistage effects, clustering, and unequal weighting, the latter two are the most critical. Other factors that can impact design effects when merging samples are sample type, post-stratification, and inclusion rate in the subpopulation. See Section 2.

The impacts of combining NAEP samples are tradeoffs between efficiency and precision. Although the efficiency of combined sample will not be as high as that of the National sample, since the sample size of a combined sample will be much larger than original National sample, the estimates in reporting results will have higher precisions. Especially, the power to measure the performance gaps of student groups of interest will increase significantly. See Sections 3 and 4.

1. Combining NAEP Reading National and State samples

A pre-condition for merging two samples is that they are equivalent (Spencer, 1997). This means two assessments should have same goal, similar instruments, scoring based on similar rubric-related features. Moreover, the two tests should be administered and supervised under similar conditions.

Combining 1998 NAEP reading National and State samples consists of two stages: i) analysis of the equivalence between National and State samples, and ii) combining National and State samples.

i) Analysis of the equivalence between National and State samples

First, the scale scores of the NAEP State samples are linked to the scale scores of the NAEP National samples (Allen et al, 2001). Then a check of the equivalence of the National and State samples is implemented. Results show that, after linking process transformation, the indeterminacy between national and state scales are diminished. By analyzing the data of the 1998 reading assessments, the combined sample and the National sample have very close distributions for main reporting groups, such as total, gender and ethnicity. In addition, the mean scale scores for the main groups from National Sample and State samples are very close in values.

ii) Combining National and State samples

To merge the NAEP National and State samples, a set of optimized shrinkage weights was created. Let $\overline{y}_{i,N}$ and $\overline{y}_{i,S}$ are the mean estimates from subsample i of National and State samples, composite estimator for subsample i is $\overline{y}_{i,C} = \alpha_i \overline{y}_{i,N} + (1 - \alpha_i)\overline{y}_{i,S}$. The calculation of optimized weights varied for subsamples of interest. The set of shrinkage weights allows mean statistics to have minimum variance estimates (Qian & Spencer, 1993; Cohen & Spencer, 1991)

2. Measures of efficiency and precision for combined sample

To measure efficiency of sampling, Kish (1965) defined *design effect* (DEFF) as a ratio of the variance of a statistic from complex samples over the variance of the statistic from simple random samples. It is also a useful tool to analyze the efficiencies of the domains in combined samples. Likewise, *relative precision* can also be used to measure precision of estimates from combined sample. It is defined as a ratio of the variance from the combined sample over the variance from the National sample (Cochran, 1977).

Several statistical factors will influence relative precision and design effects in educational assessments. They are factors of stratification, multistage effects, clustering, and unequal weighting. The last two are the most critical. Other factors that can impact precision and efficiency when merging samples are sample type, post-stratification, and inclusion rate in the subpopulation (Spencer and Liu, 1998).

The *clustering effects* are generally the dominant cause of relative precision and design effects, which can be approximated by

$$1+(\overline{M}-1)\rho$$

for mean estimates, where $\overline{M}$ is the average cluster size and ρ is the intracluster correlation (Cochran, 1977, 209). Therefore, a large cluster size or large intracluster correlation will inflate the clustering effects.

The *effects of unequal weighting* can be expressed by coefficient of variation of the mean of weights $\overline{w}$:

$$deff_W = 1 + CV_W^2$$

where CV_w is coefficient of variation of weights (Kish, 1992; Cochran, 1977). This formula assumes that the inclusion probabilities are unrelated to the measurements of interest. The effects of unequal weighting are determined by the variation of weights across primary sampling units (PSU). The effects of unequal weighting are usually stable, ranging from 1.2 to 1.3. The results of the combined samples are consistent with the findings in most of the NAEP assessments.

3. **Results for the NAEP combined samples**

Some of the findings in the analysis of the 1998 NAEP reading combined samples are in Tables 1 and 2.

i) Table 1 shows the design effects for 1998 Reading Assessment, Grade 8. By applying poststratified weights, the average design effect for non-SD/LEP student group (A2 or A3 on the Table) is 17.6. The average design effect for the combined sample (Total) is 16.5; and that for the SD/LEP student group (B2 on the Table) is 5.1. The design effects for non-SD/LEP students are the largest among three types of students.

The large design effects for non-SD/LEP students are largely due to their large clustering effects. First, cluster sizes for non-SD/LEP students are usually large. Second, compared with the SD/LEP students, the non-SD/LEP students are relatively homogenous in scale scores. The high homogeneity in clusters implies a large intracluster correlation. Everything else being equal, the large clustering effects, due to a large intracluster correlation, will boost the design effects in the reporting samples.

ii) Table 2 gives out the relative precision for total sample and for the main groups of interest. The calculations are based on design effects and sample sizes of different groups. On average, the variances of means for the combined sample will be around 39% of those for the National sample. The variances for non-SD/LEP students in the combined sample will be around 41% of those for the National sample. And those for SD/LEP students in combined sample will be about 24.7% of those for the National sample.

iii) Post-stratification of weights also reduces the design effects. The results in the Tables show that the variances of estimates applying poststratified weights tend to be smaller than those applying non-poststratification weights.

4. Conclusions

The results in i) of Section 3 show that, if the combined sample and the National sample are at same level of sample sizes, the combined sample will have a lower efficiency than the National sample. However, the actual sample size for the combined sample is much larger.

The analysis of relative precision in ii) of Section 3 show us that, although the design effects for the combined samples were large, the estimates will have smaller variances than those obtained from the National sample or State samples because the combined sample size is almost ten times as large.

For 2002 and beyond, NAEP will use combined sample to report assessment results. This study has provided a preview of the effects of using combined samples in NAEP assessments. Combined sample will increase the power to measure the performance gaps in study.

The methodology in this study provides a basis for analyzing the efficacy of merging NAEP National and State samples. The analysis of design effects will help researchers to optimize their design, minimizing cost for a specified precision.

References

Allen, N. et al. (2001). *The 1998 NAEP Technical Report*. Washington DC: National Center for Education Statistics.

Cochran, W. (1977). *Sampling Techniques,* 3rd ed. New York: John Wiley & Sons.

Cohen, T. & Spencer, B., (1991). Shrinkage Weights for Unequal Probability Samples. *Proceedings of the Section on Survey Research Methods*, 625-630.

Kish, L. (1990). "Weighting: Why, When and How?" *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 121-130.

Kish, L. (1965). *Survey Sampling,* New York: John Wiley & Sons.

Qian, J. & Spencer, B. (1993). "Optimally Weighted Means in Stratified Sampling," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 863-866.

Spencer, B. & Liu, X. (1998). Standard Errors of NAEP Statistics. *Research Paper,* Northwestern University, Evanston.

Table 1. Design Effects for Combined Reporting Samples and Subsamples For 1998 NAEP Grade 8 Reading Assessment

	Poststratified				*Not Poststratified*			
Subsamples	Total	A2*	A3*	B2*	Total	A2*	A3*	B2*
Main Reporting Groups								
Total	19.99	26.80	17.98	5.87	32.46	35.13	19.72	6.08
Male	12.50	16.11	13.84	4.71	17.69	20.43	14.28	4.59
Female	13.93	15.82	8.95	5.34	21.25	19.64	10.49	5.59
White	23.61	21.11	14.20	6.58	26.17	23.25	14.16	8.19
Black	6.51	6.77	11.14	2.37	7.06	7.33	9.39	2.16
Hispanic	11.04	7.54	7.16	5.51	10.69	7.68	7.28	5.01
Large City	14.29	12.59	24.01	12.91	14.93	13.94	21.67	11.51
Mid Size City	35.45	63.51	13.41	5.82	41.68	71.03	14.58	4.64
UF Large City	12.97	9.63	18.35	5.95	11.77	10.11	17.56	5.82
UF Mid City	15.25	16.33	25.12	2.14	15.66	15.44	25.66	1.91
Large Town	16.41	6.64	22.29	2.14	16.60	6.64	21.75	2.62
Small Town	16.75	8.77	11.27	2.13	17.27	9.57	10.83	1.82
Rural Area	36.78	44.71	15.20	11.55	44.47	49.20	16.17	12.62

Table 2. The Relative Precision for Combined Reporting Samples and Subsamples For 1998 NAEP Grade 8 Reading Assessment (in percent)

	Poststratified				*Not Poststratified*			
Subsamples	Total	A2*	A3*	B2*	Total	A2*	A3*	B2*
Main Reporting Groups								
Total	41.2	41.6	37.5	17.1	67.7	54.5	41.1	17.7
Male	37.0	37.2	40.9	20.3	52.3	47.2	42.2	19.8
Female	48.5	48.6	33.9	20.1	80.4	60.3	39.7	21.0
White	56.7	56.8	39.8	32.7	73.3	62.6	39.7	40.7
Black	27.9	53.2	27.9	13.6	17.7	57.6	23.5	12.4
Hispanic	16.7	15.9	35.4	25.4	52.9	16.2	36.0	23.1
Large City	15.3	13.6	58.7	70.9	36.5	15.1	53.0	63.2
Mid Size City	63.9	91.1	46.4	26.7	N/A**	101.9	50.4	21.3
UF Large City	21.9	18.2	39.3	31.9	25.2	19.2	37.6	31.2
UF Mid City	45.4	70.8	75.4	20.0	47.0	67.0	77.1	17.8
Large Town	8.2	N/A**	22.5	N/A**	16.7	N/A**	21.9	N/A**
Small Town	86.0	32.1	30.7	17.6	47.1	35.0	29.5	15.1
Rural Area	76.1	75.7	41.9	38.5	122.4	83.2	44.5	42.1

* A2 and A3 are two subamples of non-SD/LEP students in reporting; B2 is a subample of SD/LEP students without providing accommodations in tests.

** N/A: The nature of the sample does not allow accurate calculation of the statistic of relative precision.