Statistical Power Analysis for NAEP Combined National and State Samples

Jiahe Qian, Bruce Kaplan and Vincent Weng, Educational Testing Service
Lloyd Hicks, WESTAT, Inc.
Jiahe Qian, Rosedale Road, ETS MS 02-T, Princeton, NJ 08541

This National Assessment of Educational Progress (NAEP) study analyzes the statistical power of combined national and state NAEP samples. The goal of combining NAEP National and State samples is to provide an integral way to implement NAEP assessments and to reducing the burden of state assessments. One of National Assessment Governing Board (NAGB) policies is that national results should be estimated from state samples in order to reduce burden on states, increase accuracy, and save costs. Combined sample will provide more accurate estimations, especially for groups of small sizes such as SD/LEP students. The study is based on combined samples of the 1998 NAEP reading and 2000 math assessments.

Combining samples from different sources has previously been examined in the setting of testing. Many scholars (Boruch & Terhanian, 1999; Johnson, 1998) studied the challenging task of linking test scores from different available standardized tests so that these scores could be compared to each other and to the NAEP assessments. Some studies successfully merged data from different frames. To improve the estimation of Armed Forces Qualifying Test Score distributions in counties and battalion regions, Spencer, Nordmoe, Qian, and Haberman (1991) used combined samples from the High School and Beyond Study and the National Longitudinal Study in Analysis of Aptitude Score Distribution.

The study will introduce the process of merging the National and State samples. We focus on the analysis of the statistical power of combined samples, including studying the effective sample sizes and the necessary effective sample sizes for statistical power of tests. We also examine the changes of the significant tests in combined samples, investigate the relative precision of estimated variances and design effects.

1. Combining NAEP Reading National and State samples

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

A pre-condition for combining two samples is that samples are equivalent (Spencer, 1997). This means two assessments should have same goal, similar instruments, and scoring based on similar rubric-related features. Moreover, the two tests in merged samples should be administered and supervised under similar conditions. Feuer and Holland (1998), based on item response theory, studied the factors that cause two tests not to be parallel.

A quality control (QC) procedure was applied to check data and weights before implementing the formal analysis. The procedure included checking the distributions of the weights for the total population and subpopulations, such as race, gender, school types (i.e. public vs. private), SD/LEP, and NAEP region. For both 1998 and 2000 assessments, NAEP employed a "split sample" design. In half of the schools, students with disabilities and LEP students were allowed to use the testing accommodations they normally receive in state and district testing. In the other half, no accommodations were permitted. The two samples in the "split sample" design are called S2 and S3. The subsamples of S2 are A2, B2 and C2; for S3, the subsamples are A3, B3 and C3. The subsamples of A2 and A3 consist of non-SD/LEP students. The subsample of B2 contains SD/LEP students without providing accommodations during the assessment, whereas the subsample of B3 contains SD/LEP students, who were offered accommodations during the assessment. The subsamples of C2 and C3 are composed of SD/LEP students excluded from the assessment. For more detailed information about accommodations, see the NAEP 1998 Technical Report (Allen, Donoghue, & Schoeps, 2001) and forthcoming 2000 documentation on web site. For 2002 NAEP assessments and beyond, all the samples collected will permit accommodations.

For reporting purposes, two reporting sample types were formed to maximize the use of the collected data: **R2** and **R3**. The reporting sample type R2 refers to the samples where accommodations were not permitted; the reporting sample type R3 refers to the samples where accommodations were permitted. Specifically, R2 consists of subsamples of A2, A3, B2, and C2, and R3 consists of A2, A3, B3, and C3, appropriately weighted. Since the assessed students in R2 samples were used to report the results in the national Report Card for the NAEP 1998 and 2000 assessments, R2 is called reporting sample in this report, and R3 samples is called accommodated reporting sample.

The study included comparisons of estimates between 4th grade (G4) and 8th grade (G8) samples

with accommodation and without accommodation. The six combined samples used in the analyses are:

- 1998 G4 reading R2 sample (reporting),
- 1998 G4 reading R3 sample (accommodated),
- 1998 G8 reading R2 sample (reporting),
- 1998 G8 reading R3 sample (accommodated),
- 2000 G8 math R2 sample (reporting),
- 2000 G8 math R3 sample (accommodated).

For each sample, Westat produced two sets of weights: nonpoststratified weights and poststratified weights. The description of poststratification of weights can be found in see the NAEP 1998 Technical Report.

The national 8th grade 2000 NAEP sample was part of a pseudo-integrated design with the 8th grade State NAEP sample. The purpose of the integrated design was to allow State and national NAEP samples to be combined after data collection to produce a larger sample that would yield somewhat more precise State estimates and substantially more precise national estimates. The national 8th grade public school sample used a three-stage probability sample design, modeled after the State NAEP sample. The first-stage of selection was the sampling of schools; the second-stage was the assignment of session type and sample type to schools; and the third-stage was student sampling. The session type refers to assessment subjects (such as mathematics), while the sample type (S2 and S3) refers to the type of administration rule (accommodations offered or not). In the sampling of schools, both national and State designs explicitly stratified by state. The 2000 combined sample is a replication of the design planned for the 2002 assessments. All schools can be used for the State estimates, and all schools in the fifty states and the District of Columbia can be used for national estimates.

On the other hand, the national NAEP 1998 sample was designed independent of the State NAEP, with no intentions of combining, and had four stages of sampling. The most important difference from the 2000 national design is that its first-stage of sampling was the selection of counties or groups of counties, known as primary sampling units (PSUs). PSUs were explicitly stratified by region and metropolitan status, and one PSU was selected from each stratum. Consequently, the 1998 combined data is more limited than 2000 combined data. The 1998 combined samples will provide more precise estimates than the national data alone, nationally as well as by region, for achievement scale scores for each subject.

As results of these changes in the 2000 national sampling design, the 2000 combined samples provide accurate estimates at both national and state levels. The analysis of the equivalence between the combined sample and original samples, especially state samples, show that the new design offered consistent estimates at state levels.

ii) Combining National and State samples

To merge the NAEP National and State samples, a set of optimized shrinkage weights was created. For 1998, the calculation of the optimized weights varied for assessments (reading and writing) and was based on effective samples for the average proficiency scores (using first set of plausible values) calculated at the region level. For 2000, the composite factors or shrinkage weights were based on actual sample sizes at the state level, under the assumption that the national and State design effects are the same. The set of shrinkage weights allows mean statistics to have minimum variance estimates (Qian & Spencer, 1993; Cohen & Spencer, 1991)

2. Effects of Combined Samples on Significance Tests

The analysis in this Section consists two parts: checking the effects of combined samples on significance tests and the effects on efficiency and precision.

One concern about the moving from the national to the combined sample is the effects on increasing the power of significance tests. Since combined samples have considerable increase in the effective sample size, the sampling variability of estimates will be reduced. In the next section, we will see that the standard errors will be about seventy percent of previous ones. As a result, many differences that were not significant in previous assessments may be flagged as significant. In general, there will be more results flagged as significant, including within-year and trend results, requiring interpretation in reporting.

The changes of within-year tests and trend tests are summarized in Tables 1 and 2 separately. For within-year comparisons for 2000 math R2 combined samples, 49% of the comparisons that are not significant in tests will become significant. For 1998 G4 and G8 Reading R2 combined samples, such changes will be about 40% and 35% separately. Few tests change from significant to not significant for the combined samples in these analyses. For R3 combined samples, the results are similar. In general, more tests will become significant when using combined samples; very few comparisons will changed from significant to not significant.

The tendency of the changes of trend tests is not as prominent as those of within-year comparisons. (See Table 1.) There are more tests that change from significant to not significant. This may be explained by the fact that the data for the previous assessments exist only for the national sample. The standard errors of differences are not reduced as noticeably as those of the within-year, where both estimates in a comparison are from the combined samples. Table 2 summarizes the

results of tests of within year. In general, there will be more significance in statistical tests when using combined samples because sample sizes are increased and scores are consistent for combined samples. For mean scale scores, the increases in significance are 7 and 5 percent for R2 and R3 2000 Math samples respectively and, 8 and 11 percent for R2 and R3 1998 Reading samples.

For the results of tests of trend comparisons, similar to the within-year tests, there are more significant statistical tests when using combined samples. However, there are some exceptions (changes to not significant) although standard errors are generally smaller when using combined samples. The proportion of significance in tests may decrease for some characteristics but increase for others. For example, in the 2000 Math samples, there were two tests comparing mean scores for private schools and one test comparing scores related to the school lunch program that became not significant, while a comparison of black students becomes significant.

In all the analyses, there are no tests that changed directions; for example, the mean of first group is significantly smaller than that of second group in one sample but it becomes larger than that of second group in another sample.

3. *Effects of combined sample on efficiency and precision*

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 of variances*** can also be used to measure precision of variance estimates from combined samples. It is defined as a ratio of the variance from the combined sample over the variance from the National sample (Cochran, 1977). In comparison of data from different surveys, relative precision of variances is a suitable and standard measure to evaluate the accuracy of estimation for combined samples.

Several statistical factors will influence relative precision and design effects in educational assessments. They are stratification, multistage effects, clustering, and unequal weighting. The last two are the most critical. Other factors that can affect precision and efficiency when combining samples are sample type (accommodation rules), poststratification, and inclusion rate in the subpopulations (Spencer & Liu, 1998).

The effect of combining NAEP samples is a compromise between efficiency and precision. For more detailed descriptions of the findings of the efficiency and precision, see analysis done by Qian and Kaplan (2001). In general, because of the large clustering effects, the design effects for combined samples are relative large. Hence, at the same level of sample sizes, the combined sample will have a lower efficiency than the National sample. Although the design effects for the combined samples were large, since the combined sample size is almost ten times as large, the estimates will still have smaller variances than those obtained from the National sample or State samples. Relative precision for estimates derived from poststratified weights is smaller than those with non-poststratified weights as expected.

4. *Statistical Power of Combined Samples*

In the analysis of the statistical power of complex samples, the effective sample size and effect size play key roles. The effective sample size bridges the complex samples and simple random samples; and the effect size is used to find out the necessary sample size for given statistical power in test.

4.1 Effective Sample Sizes for Combined Samples

The effective sample size is defined by dividing a sample size by the design effect, which is equivalent to the sample size of a simple random sample. Table 3 shows the effective sample sizes for NAEP 1998 reading samples.

Since the design effects of combined samples are relative large, the increases of effective sample sizes from national samples are not proportional to the increases of sample sizes from national samples. However, the effective sample sizes of combined samples are still larger than the sample sizes of their corresponding national samples. For example, the effective sample size of the combined sample of 1998 G4 reading assessment is 8,286 and the effective sample size of its national sample is 2,436. It suggests that the statistical power of the combined samples be higher than that of the national samples.

Table 3. Effective Sample Sizes for NAEP 1998 Reading Samples

		Original Sample Size		Effective Sample Size	
	Type	National	Combined	National	Combined
G4	R2	7,672	114,826	2,436	8,267
G4	R3	7,812	115,592	1,873	7,635
G8	R2	11,051	102,257	2,085	5,115
G8	R3	11,193	103,082	2,252	5,941

4.2 Necessary Effective Sample Sizes for Comparisons

To study the power of the test of a comparison, we assess the necessary samples sizes needed for different samples. The sample sizes are all measured by the effective sample sizes needed. The alpha level of the tests is set at 0.05 and 0.01 and power is set at 0.80 and 0.90. Instead of drawing power curve, this process uses

the concept of effect sizes in determining the necessary sample sizes.

The effect size for comparison of means (Cohen, 1988) is defined as $\gamma = |\Delta| / S.D.$, where $S.D.$= $\sqrt{(n_1 S_1^2 + n_2 S_2^2)/(n_1 + n_2)}$ and $|\Delta|$ is the difference of two means. For other statistics, the analysis of effect sizes would be different. For example, the suitable measure of effect size for comparison of proportions is H index, which is obtained by arcsine transformation of proportions.

In calculation of necessary sample sizes, we assume that the sizes of two groups in comparison are equal. The relationship of sample sizes (n), effect size (γ), and effect size in standard error units (δ) is $\delta = \gamma \cdot \sqrt{n/2}$. By the table of "$\delta$ as a Function of Significance Criterion and Power" (Cohen, 1970, 824), we can calculate the effective sample sizes. For a comparison with alpha level of 5%, δ = 2.80 if power is set at 0.80, and δ = 3.24 if power is set at 0.90. For a comparison with alpha level of 1%, δ = 3.42 if power is set at 0.80, and δ = 3.86 if power is set at 0.90.

The necessary effective sample sizes for comparisons are listed in Table 4. The effect sizes in calculation are estimated from 2000 math national and combined samples. In comparison of male versus female scale scores, the sample sizes, of each gender group, for national and combined samples are 3,912 and 4,782 separately, for a test alpha level at 1% and its power set at 0.80. The total samples needed are 7,824 and 9,564. By Table 3, the national sample is not sufficient enough to obtain the statistical power, whereas the combined sample is large enough. For other comparisons, especially for small groups in comparisons, we have similar conclusions. Clearly, the combined samples provide higher power in tests than the national samples.

5. Conclusions

The findings of the study show that the combined samples will provide effective score measures, either means or achievement level percentages, same as those from the National samples. The effect of combining NAEP samples is a tradeoff between efficiency and precision. The efficiency of the combined sample will be lower than that of the National sample, but the estimates of combined samples will have higher precision. The standard errors of the scores measured would be smaller resulting in more statistically significant results in statistical tests of comparisons. Moreover, the effective sample sizes for combined samples are larger than the effective sample sizes of original national samples. Therefore, combined samples increase the statistical power to measure the performance gaps in study.

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.

References

Allen, N., Donoghue, J., & Schoeps, T. (2001). *The NAEP 1998 technical report* (NCES 2001-509). Washington DC: National Center for Education Statistics.

Boruch, R., & Terhanian, G. (1999, April). *Putting studies, surveys, and data sets together: Linking NCES surveys to one another and to data sets from other sources.* Paper presented at the annual meeting of AERA , Montreal, Canada.

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Lawrence Erbaum Assoc. Publishers.

Cohen, J. (1970). *Approximate Power and Sample Size determination for Common One-Sample and Two Sample Hypothesis Tests*. Educational and Psychological Measurement, 30, 811-831.

Cohen, T., & Spencer, B. (1991). Shrinkage weights for unequal probability samples. *1991 Proceedings of the American Statistical Association, Survey Research Methods Section*, 625–630.

Feuer, M. J., Holland, P., Green, B. F., Bertenthal, M. W., & Hemphill, F. (Eds.). (1998). *Uncommon measures: Equivalence and linkage among educational tests*. Washington DC: National Academy of Science.

Johnson, E., et al. (1987). *Weighting Procedures in Implementing the New Design: The NAEP 1983–1984. Technical Report*. Princeton, N.J.: National Assessment of Educational Progress.

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

Qian, J., & Spencer, B. (1993). Optimally weighted means in stratified sampling. *1993 Proceedings of the American Statistical Association, Survey Research Methods Section*, 863–866.

Qian, J., & Kaplan, B. (2001). Analysis of Design Effects for NAEP Combined Samples. *2001 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM]*, Alexandria, VA: American Statistical Association.

Spencer, B., Nordmoe, E., Qian, J., & Haberman, S. (1991). *Aptitude Score Distribution Study—Phase 1: Sampling Merging and Direct Estimation of Armed Forces Qualifying Test Score Distributions in Counties and Battalion Regions.* Chicago: NORC.

Spencer, B. (1996). *Combining State and National NAEP.* Paper presented at the conference of the National Assessment Governing Board. Washington, DC.

Table 1. Changes in significant tests (within year & trend)
For NAEP 1998 Reading and 2000 Math Assessments

Measure and Sample	1998 Reading, G4		1998 Reading, G8		2000 Math, G8	
Mean & Achiev. Level	Proportion of changes: not sig.to sig.	Proportion of changes: sig.to not sig.	Proportion of changes: not sig.to sig.	Proportion of changes: sig.to not sig.	Proportion of changes: not sig.to sig.	Proportion of changes: sig.to not sig.
***R2** (within year)*						
NS vs. PS CS*	0.43	0.01	0.36	0.00	0.49	0.03
NS vs. Non-PS CS	0.37	0.01	0.34	0.01	0.49	0.04
Non-PS CS vs. PS CS	0.09	0.00	0.10	0.02	0.07	0.01
***R3** (within year)*						
NS vs. PS CS	0.43	0.00	0.36	0.00	0.42	0.02
NS vs. Non-PS CS	0.41	0.00	0.27	0.00	0.40	0.02
Non-PS CS vs. PS CS	0.06	0.01	0.12	0.00	0.07	0.01
***R2** (trend)*						
NS vs. PS CS	0.08	0.00	0.12	0.19	0.03	0.14
NS vs. Non-PS CS	0.05	0.13	0.14	0.11	0.03	0.36
Non-PS CS vs. PS CS	0.03	0.00	0.00	0.09	0.07	0.33
***R3** (trend)*						
NS vs. PS CS	N/A	N/A	N/A	N/A	0.00	0.32
NS vs. Non-PS CS	N/A	N/A	N/A	N/A	0.01	0.68
Non-PS CS vs. PS CS	N/A	N/A	N/A	N/A	0.11	0.33

* NS=National sample; CS=Combined Sample; PS=poststratified. Note that no R3 samples were collected in 1994 NAEP assessments, so no trend comparisons can be computed for the R3 samples.

Table 2. Proportion of Significance in Tests (within year)
For NAEP 2000 Math and 1998 Reading, Reporting Sample

Measure and Sample	1998 Reading, G4		1998 Reading, G8		2000 Math, G8	
Mean & Achiev.Level	Significant	Not sig.	Significant	Not sig.	Significant	Not sig.
R2						
NS	0.68	0.32	0.62	0.38	0.70	0.30
PS CS	0.81	0.19	0.76	0.24	0.83	0.17
Non-PS CS	0.79	0.21	0.74	0.26	0.82	0.18
R3						
NS	0.64	0.36	0.61	0.39	0.72	0.28
PS CS	0.80	0.20	0.75	0.25	0.82	0.18
Non-PS CS	0.79	0.21	0.72	0.28	0.82	0.18

Table 4. Necessary Effective Sample Sizes for Each Group in Comparisons
for NAEP 2000 Math Samples (Poststratified) *

		Alpha=0.05				**Alpha=0.01**			
Comparisons		**Power=0.80**		**Power=0.90**		**Power=0.80**		**Power=0.90**	
		Nat.	Comb.	Nat.	Comb.	Nat.	Comb.	Nat.	Comb.
Male	Female	3,202	3,915	4,183	5,113	3,912	4,782	4,983	6,091
Black	Hispanic	676	323	883	421	826	394	1,052	502
Black	Amer Ind	304	260	397	339	371	317	473	404
Post HS	Grad Col	357	280	466	366	436	342	555	436
Public	Nonpublc	147	171	192	223	179	208	228	265
SE	WEST	600	485	784	633	733	592	934	754
Cent.city	Urb. fring	171	165	224	216	209	202	267	257
Cent.city	Rural	406	430	530	562	496	526	631	670

* Assume that the sizes of two groups in comparison are close. The effect sizes in calculation are estimated from 2000 math national and combined samples.