One approach to detecting the invariance of proficiency standards over time

Jiahe Qian[1]
Educational Testing Service, Princeton, NJ, USA[2]

This study explores the use of a mapping technique to test the invariance of proficiency standards over time for state performance tests. The first step involved mapping the state proficiency standards onto the National Assessment of Educational Progress (NAEP) scale. During the second step, no attempt was made to determine a direct deviation in proficiency standards. Instead, this step involved testing the invariance of the NAEP equivalents of the state standards over time. The basis of the mapping technique is an enhanced method that was originally designed for comparing performance standards for public school students set by different states when the state tests are comparable. This approach can also be used to detect score inflation over time for state tests.

[1] The author thanks Shelby Haberman, Henry Braun, Sandip Sinharay, and William Monaghan for their suggestions and comments. The author is particularly grateful to Daniel Eignor for his help in clarifying the psychometric meaning of the statistical techniques. The author would also like to thank Bruce Kaplan, Sailesh Vezzu, and Xiaoke Bi for their computational assistance and Kim Fryer for editorial assistance. The opinions expressed herein are solely those of the author and do not necessarily represent those of the Educational Testing Service. This work was supported by the National Center for Education Statistics, Contract # ED-02-CO-0023.

[2] The opinions expressed herein are those of the author and do not necessarily represent those of Educational Testing Service.

INTRODUCTION

In the United States, the stability of state education performance test standards has recently become a concern. The reason is that, under the No Child Left Behind Act (NCLB), each state can select its own tests and set its own proficiency standards for reading and mathematics, thereby determining its standing with respect to the national requirements of adequate yearly progress (American Federation of Teachers, 2006). The study presented here was designed to develop an approach to test the level of invariance in proficiency standards over time for state tests or analogous assessments.

Proficiency standards—specific levels of mastery of knowledge and skills in education—are usually anchored by cutoff points on a test scale. These cutoff points classify student performance into several achievement categories, such as *basic, proficient*, and *advanced*. State tests usually use an equating process to ensure numerical cutoff points reflect proficiency standards, provided that no substantial changes occur in the assessment. However, over time, proficiency standards can deviate from the achievement levels on the original scale on which they were established, such that each cutoff point no longer anchors to the same ability level. This phenomenon is called *deviation in proficiency standards* (DPS) or *proficiency standard deviation*. Many researchers have found DPS in performance assessments when they compare these assessments with other stable assessments such as the National Assessment of Educational Progress (NAEP) (Cannell, 1987; Grissmer, Flanagan, Kawata, & Williamson, 2000; Klein, Hamilton, McCaffrey, & Stecher, 2000; Neill & the Staff of FairTest, 1997; Smith, 1991). Another concept, *scale drift*, also relates to the stability of a test scale. However, as Angoff (1984) notes, scale drift usually relates to inadequate equating of a new form of a test to "one or more of the existing forms for which conversions to the reference scale (i.e., the reporting scale) are already available" (p. *viii*).

DPS can be caused by many factors, such as score inflation, scale drift, alteration of the test instrument, changes in assessment format, reform of the subject framework, content modification, or differential performance gains (Koretz, 2007; Madaus, 1988b). However, if other factors are not present, DPS can serve as an indicator of score inflation, meaning that students gain higher test scores than before at each given level of academic achievement (Arenson, 2004; Koretz, 1988; Linn, 2000; Potter, 1979). The section of this paper headed "Application to Empirical Data" discusses a procedure for detecting score inflation by proxy: detect the presence of DPS, a necesssary condition for score inflation.

The approach developed in this study to test the invariance of proficiency standards is based on an enhanced mapping technique that was originally designed for comparing performance standards that different states set for public school students when the tests are comparable (Braun & Qian, 2007a). Testing whether the proficiency standard deviates from its original scale by simply observing the changes of scores on a test itself is difficult. However, it becomes feasible when a test with potential DPS is compared with another test that has no DPS. For example, many educators have compared state tests with the NAEP assessments. They found that test score improvements shown on state tests used for high-stakes decisions tend not to be corroborated by score improvements on the NAEP (Haney, 2002; Linn, Graue, & Sanders, 1990). This study employs an analogous strategy to measure the invariance of proficiency standards. This strategy transforms the scale of state tests to the well-established NAEP scale and then uses the NAEP scale as a benchmark to detect whether the proficiency standards of state tests are invariant over time.

To implement the approach developed by this study, the state proficiency standards were first mapped onto the NAEP scale. (These mapped proficiency standards of state tests are called the *NAEP equivalents to the state standards* or *NAEP equivalents*.) Next, a related solution—the invariance of the NAEP equivalents over time—was examined. This was done in preference to effort focused on directly detecting invariance of proficiency standards. The mapping makes the comparison effective because, as a benchmark, NAEP is generally regarded as meeting high standards with respect to test design, test content, and psychometric quality. In addition, NAEP is the only nationally standardized test that is administered in a uniform and stable manner across states. Also, NAEP scores are not influenced by factors such as grade inflation. (For a general introduction to NAEP, see Jones & Olkin, 2004.)

Although the literature demonstrates, for a number of reasons, that linking state tests to NAEP assessments at the student level does not result in an appropriate or a valid linking (Feuer, Holland, Green, Bertenthal, & Hemphill, 1998; Koretz, Bertenthal, & Green, 1999), studies show that mapping the proficiency standards on state tests to NAEP equivalents is valid (Braun & Qian, 2007b; McLaughlin & Bandeira de Mello, 2003). Because the schools in the sample used in this study take both NAEP and state tests, there is an overlapping of student populations between the two assessments. Moreover, the test instruments used for NAEP and the state tests are similar but not the same. These features ensure that the mapping procedure provides a valid assessment of state standards in terms of stability. Therefore, most of the heterogeneity across states in the NAEP equivalents to the state standards can be attributed to differences in the stringency of proficiency standards set by the states.

If a significant change in the NAEP equivalents is found over time, then it is likely that the proficiency standards of the state test have deviated from its original scale. In this paper, it is suggested that parties interested in confirming the causes of significant DPS, especially serious claims such as score inflation, form a committee, with members consisting of test experts and subject-matter specialists able to judge the reasons for an observed change.

The next section of this paper describes the estimation method that is used to map the proficiency standards of state tests. This section also introduces some properties of the mapped proficiency standards over time. The following section introduces the data used in the study, namely the 2003 and 2005 fourth- and eighth-grade state tests of reading and mathematics. It also presents the empirical results from testing the invariance of state standards. The penultimate section documents the application of the approach used to detect score inflation for state tests. The final section offers a summary and some conclusions.

METHODOLOGY

A. Outline of the Methodology for Mapping State Standards to the NAEP Scale

As described in Braun and Qian (2007a), the mapping procedure is carried out separately for each state that participates in NAEP and is represented in the National Longitudinal School-Level State Assessment Score Database[2] (for the corresponding academic year). To make the comparisons of the NAEP equivalents over time effective, both the state tests and the NAEP assessments need to comply with standard conditions (see below).

The statistical analysis presented in this study involves the sample design of NAEP assessments, school weights, and target estimation, among other features. In NAEP, state samples are obtained through a two-stage probability sampling design. To account for the unequal probabilities of selection and to allow for adjustments for non-response, each school and each student is assigned separate sampling weights. This study applied appropriate weights when estimating the proportion of students in the state who scored above the standard. The statewide target proportion of students meeting the standard is estimated by a ratio estimator. Appendix A provides a description of the weights, the target estimation, and the variance estimation.

Let P denote the state-wide proportion of students meeting a particular standard. Let F denote the score distribution on the NAEP assessment for the state and the $(1-P)$th quantile on F be $\xi = F^{-1}(1-P)$. The estimate of the $(1-P)$th quantile, $\hat{\xi}$, can also be denoted as $\hat{\xi}_{WAM}$, where the abbreviation WAM stands for "weighted aggregate mapping." Braun and Qian (2007a) followed the steps below when mapping state standards to the NAEP scale:

1. Based on the proportions of students who meet a given state's performance standard on that state's own assessment in NAEP-sampled schools, estimate the proportion of students in the state as a whole who meet the state's standard. First, identify the schools in the state's NAEP sample and match the schools with their records in the NLSLSASD. For each school, obtain the proportion of students meeting the state standard.

   Using the school weights from the NAEP design allows one to obtain an estimate of P via a ratio estimator, $\bar{p}_w$, which is a weighted average estimate of the number of students meeting the standard over a weighted average estimate of the number of eligible students. (For a more detailed description of the weights and the ratio estimator, see Appendix A.)

2. Based on the NAEP sample of schools and students within schools, estimate the distribution of scores on the NAEP assessment for the state as a whole. This

[2] The National Longitudinal School-Level State Assessment Score Database (NLSLSASD; www.schooldata.org) is constructed and maintained by the American Institutes for Research (AIR) for the National Center for Education Statistics (NCES). Its purpose is to collect and validate data from state testing programs across the United States. It contains assessment data for approximately 80,000 public schools in the United States and is updated annually.

procedure is carried out in order to generate the results contained in the report that is issued after each NAEP assessment. Let $\hat{F}$ denote the empirical distribution of F, which can be obtained from the NAEP sample.

3. Find the point on the NAEP score scale at which the estimated proportion of students in the state scoring above that point equals the proportion of students in the state meeting the state's own performance standard. After estimating by $\bar{p}_w$ the proportion P of students meeting the state's own performance standard (defined with respect to the state test score scale) and calculating the NAEP score distribution as in Steps 1 and 2, map the performance standard to the NAEP scale by finding the point $\hat{\xi}$ on the NAEP scale that is the $(1-\bar{p}_w)$th quantile:

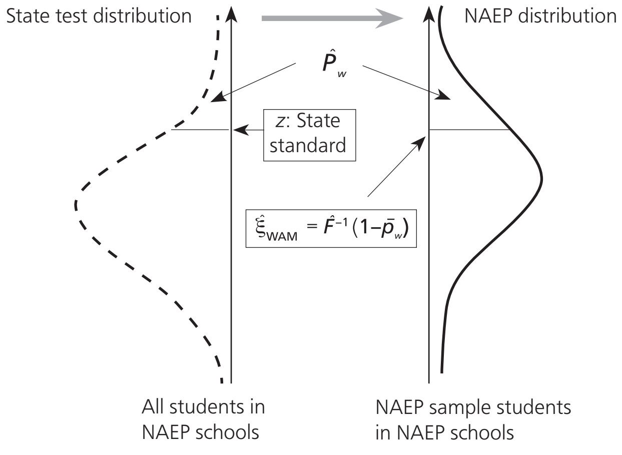
$$\hat{\xi}_{WAM} = \hat{F}^{-1}(1-\bar{p}_w). \tag{1}$$

The estimated NAEP equivalent to the state standard is taken to be $\hat{\xi}_{WAM}$, which is an estimate of ξ. If the state employs more than one standard, this procedure can be repeated for each one.

4. Compute an estimate of the variance of the estimated NAEP equivalent. Given NAEP's complex sample design and latent ability measurement, this computation is developed according to the NAEP jackknife methods used to obtain variance estimates (Allen, Donoghue, & Schoeps, 2001).

Figure 1 illustrates the mapping procedure. The dashed curve on the left-hand side represents an estimate of the state distribution of scores on the state test, based on the scores of all students in the schools selected for the state's NAEP sample. The area in the upper tail of this distribution above the state standard is an estimate of the proportion of students in the state meeting or exceeding that standard, and is denoted by $\hat{p}_w$. In practice, it is necessary to obtain only $\hat{p}_w$ from the data. The curve on the right-hand side represents the estimated distribution of NAEP scores for the state. This is the usual reported NAEP distribution that is estimated based on the performance of students in the state's NAEP sample who took the NAEP assessment.

Figure 1: The schematic of the mapping procedure


State test distribution
NAEP distribution
$\hat{P}_w$
z: State standard
$\hat{\xi}_{WAM} = \hat{F}^{-1}(1-\bar{p}_w)$
All students in NAEP schools
NAEP sample students in NAEP schools

The estimated NAEP equivalent to the state standard, $\hat{\xi}$, is the point on the NAEP scale where the corresponding upper tail area of the NAEP distribution also equals $\hat{p}_w$. For a given distribution of state test scores and a specific distribution of NAEP assessment scores, by the monotone property of equipercentile linking, a larger $\hat{p}_w$ corresponds to a lower $\hat{\xi}$ and vice versa.

B. Testing the Invariance of State Standards Over Time

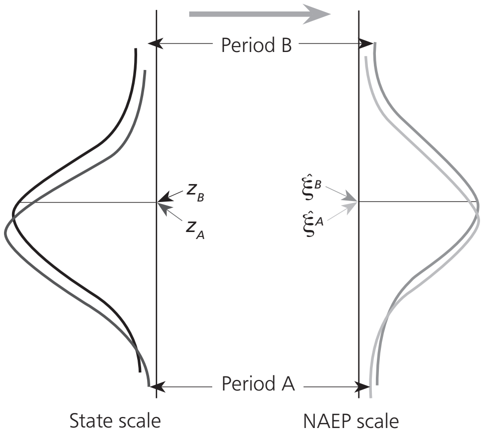
Mapping State Standards to the NAEP Scale Over Time

As I pointed out in the introductory section, the validity of the mapping methodology requires that the state test and the NAEP assessment be reasonably equivalent with respect to their test instruments, including subject frameworks, assessment format, psychometric characteristics of the tests, norms, and so on. The standard conditions involved with the procedure are then described, and the following is assumed: (a) there are no considerable changes in the state test instrument over time; (b) the state tests maintain their numerical cut points related to the standards over time (via score equating); and (c) the distributions of state test scores over time maintain the same shape and spread, but there is allowance for horizontal shifting of the distribution curve. The same assumptions are applied to the NAEP assessments. These standard conditions are reasonable even though they may appear to be stringent.

Let z_A and z_B be the state test standards of Time Point A and Time Point B, respectively. Because state tests are assumed to maintain their standards over time, $z_A = z_B$. Let ξ^A and ξ^B be the images of z_A and z_B for Time Point A and Time Point B, respectively. Their estimates are $\hat{\xi}^A$ and $\hat{\xi}^B$. The variance estimation of $\hat{\xi}^A$ or $\hat{\xi}^B$ is the same as that for $\hat{\xi}$ in Section A of the Methodology.

Let P^A be the proportion of students meeting the standard z_A for Time Point A, and let P^B be the proportion for Time Point B. The two empirical curves on the left side of Figure 2 illustrate a change between two time points, whereas $\hat{\xi}^A$ and $\hat{\xi}^B$ on the right-hand side of the figure are the results of mapping procedure.

Figure 2: The mapping procedure for a state test over two time periods



Let $P^B = P^A + \Delta P^S$, where ΔP^S is the change in the proportion of students meeting the standard in the state test. When $\Delta P^S > 0$, it means that a higher proportion of students met the standard at Time Point B. A higher proportion meeting the standard at Time Point B could occur for one of two possible reasons: there has been real progress in education or there is DPS in the testing results. If there has been progress in education, we can assume that the students will show a similar degree of progress in both the state test and the corresponding NAEP assessment.

Some Properties of the NAEP Equivalents Over Time

Let F^A and F^B denote the estimated distributions on the NAEP scale for Time Point A and Time Point B. As given in (1), the NAEP equivalent for Time Point A, the image of P^A, is the $(1-P^A)$th quantile on F^A:

$$\xi^A = F^{-1,A}(1-P^A), \tag{2}$$

and the image of P^B on F^B is

$$\xi^B = F^{-1,B}(1-P^B) = F^{-1,B}(1-(P^A + \Delta P^S)). \tag{3}$$

Let P^α be the true proportion of students whose scores are greater than the point of ξ^A in the NAEP assessment at Time Point B, that is,

$$\xi^A = F^{-1,B}(1-P^\alpha)\ . \tag{4}$$

Because of the changes in performance over time, P^α is usually not equal to P^A. Thus, $P^\alpha = P^A + \Delta P^N$, where ΔP^N is the changed proportion in NAEP above ξ^A at Time Point B.

First, assume $\Delta P^S = \Delta P^N$. Thus, for the time period, students show the same change in achievement in both the state test and the corresponding NAEP assessment. This assumption implies that $P^\alpha = P^A + \Delta P^S$. Because of (4) and

$$\xi^B = F^{-1,B}(1-(P^A + \Delta P^S)) = F^{-1,B}(1- P^\alpha), \tag{5}$$

thus $\xi^B = \xi^A$. This outcome indicates that when $\Delta P^S = \Delta P^N$, the NAEP equivalent is invariant over time. Accordingly, ξ^A can be viewed as being an *invariant equivalent*. Figure 2 illustrates how the mapping procedure for both the state test and the NAEP assessment performs for the time period in question. When the NAEP scale is used as the benchmark for comparison, invariance of NAEP equivalents over time under the standard conditions is equivalent to the invariance of state proficiency standards over time.

Second, assume $\Delta P^S > \Delta P^N$, that is, $P^A + \Delta P^S > P^A + \Delta P^N$, so the proportion of students meeting the standard on the state test is higher than that on the NAEP assessment. Because

$$\xi^A = F^{-1,B}(1- P^\alpha) = F^{-1,B}(1-(P^A + \Delta P^N)) \tag{6}$$

and the monotone property of $F^{-1,B}(\cdot)$, it follows that

$$\xi^B = F^{-1,B}(1-(P^A + \Delta P^S)) < F^{-1,B}(1-(P^A + \Delta P^N)) = \xi^A. \tag{7}$$

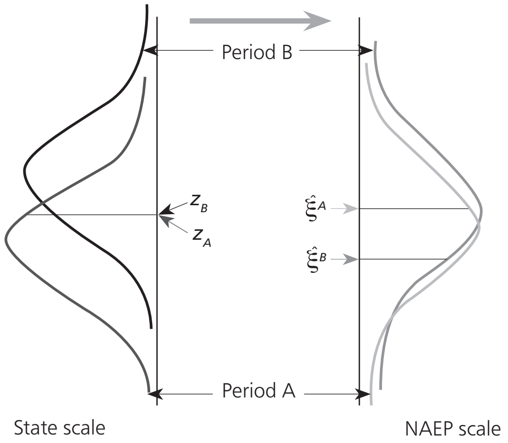
This indicates that the NAEP equivalent at Time Point B is lower than ξ^A. It shows an occurrence of DPS, a deviation in proficiency standard. Figure 3 illustrates the empirical mapping procedure indicating that the state test performs differentially from the NAEP assessment over time.

Third, assume $\Delta P^S < \Delta P^N$, that is, $P^A + \Delta P^S < P^A + \Delta P^N$, so the proportion of students meeting the standard in the state test is lower than that in the NAEP assessment. Because of (6) and the monotone property of $F^{-1,B}(\cdot)$, it follows that

$$\xi^B = F^{-1,B}(1-(P^A + \Delta P^S)) > F^{-1,B}(1-(P^A + \Delta P^N)), \qquad (8)$$

hence $\xi^B > \xi^A$. This is a trivial case, although it shows an occurrence of DPS.

Figure 3: The mapping procedure with score inflation in the state test for Period B



Test of the Invariance of NAEP Equivalents Over Time

In this study, the evaluation procedure employs both statistical significance tests and effect size criteria. For the statistical approach, the hypothesis serves as a check of the invariance of the NAEP equivalents over time under the standard conditions. The null hypothesis can be expressed as $H_o : \xi^B = \xi^A$. An equivalent hypothesis is whether the proportion of students passing ξ^B at Time Point B equals the proportion passing the invariant equivalent at Time Point B: $P^B = P^\alpha$. This study employed two significance tests in the analysis. The first test required use of a *t*-type statistic to check the difference of two proportions. The second statistic is the log-odds ratio (Haberman, 1978).

Let $n_{B.}$ be the sample size in consideration for Time Point B. Let $\hat{\xi}^B$ and $\hat{\xi}^A$ be the estimates of ξ^B and ξ^A, respectively. In Table 1, let n_{11} and n_{21} be the numbers of students whose scores are greater than $\hat{\xi}^B$ and $\hat{\xi}^A$, respectively, and n_{12} and n_{22} be the numbers of students who failed to meet the standards. Let $\hat{p}^B_w = n_{11}/n_{B.}$ be the estimate of P^B, and $\hat{p}^\alpha_w = n_{21}/n_{B.}$ be the estimate of P^α. Let $\hat{p}^B = n_{.1}/n$ and $\hat{q} = 1 - \hat{p}$.

Define the Z_c statistic as

$$Z_c = \frac{\left| \hat{p}_w^{\alpha} - \hat{p}_w^{B} \right| - 1/n_{B.}}{\sqrt{2\hat{p}\hat{q}/n_{B.}}} \ . \tag{9}$$

The term, $1/n_B$, in (9) is the Yates' correction for continuity (Yates, 1934). The log-odds ratio is defined as

$$L = \log\left(\frac{n_{11}n_{22}}{n_{12}n_{21}}\right), \tag{10}$$

and an estimate of its standard error is

$$SE\,(L) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \ . \tag{11}$$

Because the NAEP state data are collected by a two-stage sampling approach, the formulas for simple random sampling underestimate the variances employed in the test statistics. The variance estimation for complex data usually uses replicate resampling approaches (Wolter, 1985). To simplify computations and count the effects of complex sampling, the variances are estimated by multiplying a variance estimate by a design effect, which Kish (1965) introduced as a ratio of the variance of a statistic from complex samples over the variance of the statistic from simple random samples. Based on previous NAEP analyses, 2.5 is used as the approximate design effect for computation purposes. A .05 alpha level is then used in the analyses of the statistical tests.

Table 1: Number of NAEP students whose scores were greater than ξ^B and who passed ξ^A at Time Point B

	Proficiency standard		
	Pass	*Fail*	*Total*
ξ^B	n_{11}	n_{12}	$n_{B.}$
ξ^A	n_{21}	n_{22}	$n_{B.}$
Total	$n_{.1}$	$n_{.2}$	N

When the hypothesis is rejected, it means that there is a significant difference in the NAEP equivalents over time, which implies a significant DPS. It also means that the students in the state test performed differently from how they would have on the NAEP assessment. However, DPS cannot be considered equivalent to score inflation because other possible factors could cause such differences, including differential performance gains and style of classroom instruction. Only when other potential factors can be dismissed can DPS be taken as an indication of score inflation.

For practical purposes, the effect size criterion is also used to evaluate the difference of two proportions drawn from independent samples, or the differences between a single proportion and any specified hypothetical value. The effect size for comparison of proportions is called the *H* index. To provide a better scale for looking at differences on which effect sizes for proportions are comparably detectable, Cohen (1988) applied the arcsine transformation to the proportions before calculating their difference.

Let the arcsine transformation be $\varphi = 2\arcsin\sqrt{p}$. The *H* index for proportions is then defined as $H = |\varphi_1 - \varphi_2|$. To count as an intermediate effect size, the absolute value of the *H* index has to be at least 0.20 in the measuring differences of two proportions.

C. Evaluating the Test Results

When a significant DPS is detected, it is important to find the causes of this deviation in respect of proficiency standards. To assess the test results, a committee of test experts and subject-matter specialists should be assembled. This process is analogous to the review process that happens with a NAEP differential item functioning (DIF) analysis (Allen et al., 2001).

The process of judging the causes of the deviation in the NAEP equivalents over time consists of two phases. The initial phase involves executing the relevant computations and statistical tests. The second phase involves assessing the results and determining the factors that could cause the deviation in proficiency standards over time. The expert committee will need to check the standard condition assumptions, review the results of the findings, discuss the possible causes for the differences, and draw conclusions. Only if all competing potential causes are eliminated can the results be attributed to score inflation.

APPLICATION TO EMPIRICAL DATA

Data

To detect the deviation in proficiency standards over time, this study analyzed two sets of data: (a) the 2003 and 2005 NAEP mathematics and reading assessment samples for Grade 4 (G4) and Grade 8 (G8) students; and (b) the 2003 and 2005 state test samples of mathematics and reading for G4 and G8 students. The information on the proportions of students meeting state test standards for 2003 and 2005 was retrieved from the NLSLSASD. This database contains the proportions of students, by school, meeting each of the state's standards for nearly all states, beginning as early as the academic year 1994. However, it does not contain scores for individual students. The NLSLSASD typically presents, for each school, the percentage of students meeting or exceeding each achievement standard established by the state.[3]

[3] For almost all states, some schools in the NAEP school sample were either missing from the NLSLSASD or the required datum was not listed. In these cases, the number of schools available for estimation was smaller than the number of schools in the NAEP school sample. For each subject and grade combination, there were four to five jurisdictions in which the proportion of NAEP sample schools employed in the estimation was less than 0.9.

Empirical Results

The analysis involved first completing the mapping procedure described in the section of this paper headed "Outline of the Methodology for Mapping State Standards to the NAEP Scale." The report prepared by Braun and Qian (2007b) contains the tables of the estimates of the statewide proportion of proficient students, the estimated NAEP equivalents to the state standard, and the estimated standard error of the NAEP equivalent for the 2005 G4 and G8 reading and mathematics state tests, respectively. The same report also contains the results for the 2003 G4 and G8 reading and mathematics state tests, respectively. For each state, each table also displays the number of schools in the NAEP sample and the number of schools employed in the mapping. This last quantity is simply the number of schools in the NAEP sample that could be matched to the schools with usable state test performance data. The notes under each of the tables list issues concerning data in this present analysis.

Appendix B contains four figures pertaining to the ordered estimated NAEP score equivalents together with their estimated standard errors for the four subject and grade combinations (reading, Grades 4 and 8; mathematics, Grades 4 and 8). The estimated standard errors in these figures are relatively small compared to the range of the estimated NAEP score equivalents. The error bands in the figures extend plus or minus 1.96 standard errors on either side of the estimated NAEP equivalent for the state.

As shown in Figure B1, for reading at G4 in 2005, the largest estimated NAEP score equivalent of 234 (Massachusetts) is 73 points higher than the lowest one, 161 (Mississippi). The other figures, B2 to B4, also show similarly wide ranges of the estimated NAEP score equivalents. The large discrepancies in the mapped states' assessment standards make it difficult to gauge where states currently are in terms of reaching the goal delineated by the No Child Left Behind Act (Lewin, 2007) and how far they have to go to reach that goal.

For the G4 reading analysis, this study used, for comparative purposes, the data from 21 of the 25 states that had both 2005 and 2003 data. To align the state test and the NAEP reading assessments, this study dropped state data if the relevant state assessment was labeled "English/Language Arts" rather than "Reading." The study only considered and later discussed those states that showed an increase in the proportion of students meeting the state standards. The outcome of both the statistical tests and the effect size check showed significant results for two states. These two (States 1 and 2) are listed in Table 2.

Note that the names of all the states listed in Table 2 are unspecified because the possible causes for the deviation in proficiency standards have yet to be investigated. For example, the State 1 test shows a large increase in the proportions of students meeting the state's standards. In the 2005 NAEP sample for State 1, the proportion of the students who passed ξ^B is 0.71, and the proportion of those who passed ξ^A is about 0.60. The images of $\bar{p}_w^A$ on $\hat{F}^A(0.60)$ and $\bar{p}_w^B$ on $\hat{F}^B(0.71)$ show significant variation in the NAEP scale over time. This pattern indicates the presence of a significant DPS, or a deviation in state proficiency standards.

Table 2: Results of statistical testing and *H* index checking, including significance, for Grade 4 reading and mathematics

State	2005 Estimate of proportion passing $\hat{\xi}^B, \hat{p}_w^B$	2005 Estimated NAEP equivalent, $\hat{\xi}^B$	2005 Estimate of proportion passing $\hat{\xi}^A, \hat{p}_w^\alpha$	2003 Estimated NAEP equivalent, $\hat{\xi}^A$	Z_c statistic	Log-odds ratio	*H* index
			Grade 4 reading				
1	0.71	202	0.60	212	6.61	0.21	0.23
2	0.80	197	0.67	210	6.89	0.29	0.30
			Grade 8 reading				
3	0.63	244	0.52	256	5.15	0.19	0.22
4	0.82	235	0.73	247	5.24	0.23	0.22
5	0.72	245	0.63	256	5.56	0.18	0.19
6	0.30	276	0.19	285	6.02	0.27	0.26
7	0.57	254	0.43	267	6.48	0.25	0.28
			Grade 4 mathematics				
8	0.85	218	0.76	226	5.72	0.25	0.23
9	0.80	224	0.65	234	6.79	0.33	0.34
10	0.91	207	0.78	217	8.50	0.45	0.37
			Grade 8 mathematics				
11	0.61	269	0.52	278	4.35	0.18	0.20
12	0.53	276	0.44	286	4.51	0.17	0.20
13	0.74	258	0.64	268	4.68	0.20	0.22
14	0.70	266	0.53	280	8.24	0.32	0.35
15	0.65	277	0.44	293	8.82	0.37	0.42

In the G8 reading analysis, this study used, in the comparison, data from 28 of the 30 states that had both 2005 and 2003 data. Significant results relative to both the statistical testing and the effect size check emerged for five states (States 3 to 7). Table 2 displays the results for these five states.

In the G4 mathematics analysis, this study used data from 24 of the 25 states with both 2005 and 2003 data. After the first phase of the analysis, three of the states (States 8 to 10) listed in Table 2 showed significant differences in the NAEP equivalents as an outcome of the statistical testing and the effect size checks. Of the three state tests, the State 8 test showed a substantial increase in the proportions of students meeting its standards. Seventy-four percent and 85% of the students passed its standards in 2003 and 2005, respectively.

In the 2005 NAEP sample for State 8, the proportion of the students who passed its ξ^A is 0.76. The tests produced a significant outcome for the variation of the images of $\bar{p}_w^A$ on $\hat{F}^A$ (0.76) and $\bar{p}_w^B$ on $\hat{F}^B$ (0.85). This variation implies that the G4 State 8 mathematics test has a significant DPS and thus a deviation in state proficiency standards.

To confirm the cause of these changes in achievement level percentages, further investigation is needed, and final approval must be acquired from an expert committee during a second-phase analysis.

In the G8 mathematics analysis, this study used data from 25 of the 32 states with both 2005 and 2003 data. Significant results emerged for five states (States 11 to 15) with respect to the statistical testing and the effect size check.

AN APPLICATION: DETECTING SCORE INFLATION IN THE STATE TESTS

An important application of this approach is detecting score inflation in state tests. If other factors causing DPS can be excluded, a significant DPS indicates score inflation. DPS is thus a necessary condition for the demonstration of score inflation.

Over recent years, score inflation has become an increasing concern for many educators because it compromises efforts to improve education and accountability in assessments (Bromley, Crow, & Gibson, 1973; Hambleton et al., 1995; Rosovsky & Hartley, 2002; Shepard, 1988). Score inflation can be tied to a variety of situations. For non-linked or poorly equated tests, lack of adequate equating can result in what might be considered to be grade inflation. But even if the scale of a test is well linked or equated, score inflation can still be present.

A typical situation occurs when classroom instruction is test-driven or when students are focused on learning content specific to the questions asked in a standardized test. Because students at different achievement levels know the content of questions because they have memorized the same answers, the resulting scores will not necessarily indicate the real academic level of individual students. In particular, students at a lower proficiency level often achieve test scores that are higher than their relative aptitude in such environments would predict (Haladyna, Nolan, & Haas, 1991; Madaus, 1988a; Phelps, 2005). Such situations result in assessments that fail to adequately measure student levels of achievement; even efforts to align tests closely with curricular standards are insufficient to guard against this sort of score inflation (Koretz, 2005).

The principle applied to test score inflation is to check whether score improvements on NAEP corroborate score improvements on the state tests. The stability of NAEP scales is thus the basis of such comparisons. If a DPS is detected, an expert panel should then be asked to determine if the cause of the deviation is likely due to score inflation.

Of the two cases of DPS discussed in the section above on testing the invariance of state standards over time, only one gave an indicator of score inflation. When $\Delta P^S > \Delta P^N$, it implied that the proportion of students meeting the standard on the state test was higher than that on the NAEP assessment. The NAEP equivalent at Time Point B was lower than ξ^A, that is, $\xi^B < \xi^A$. This case of significant DPS accordingly provided a scenario for possible score inflation. In the second case, where $\Delta P^S < \Delta P^N$, the implication was $\xi^B > \xi^A$. This case of significant DPS was not an indicator of score inflation. It may have occurred because of failure to satisfy standard conditions or because of a change in testing conditions.

To reemphasize, in order for analysts to formally claim score inflation, they must ensure that the causes of DPS are evaluated by an expert committee, and they must discuss potential factors other than score inflation. Although analysis of the 2005 and 2003 G4 and G8 reading and mathematics data presented in Table 2 demonstrated significant DPS, specific causes of DPS, including possible score inflation, were not determined because these results had not been reviewed by an expert committee.

Finally, it is possible that the changes in NAEP equivalents over time were caused by a combination of factors: they may have been partly due to modification of the item formats and test structures and partly due to score inflation. Resolving this situation and drawing conclusions will necessitate the collection of additional data in further studies.

CONCLUSIONS

The study presented in this paper developed an approach for testing the invariance of state proficiency standards over time for state tests or other analogous assessments. This approach is based on the methodology originally developed for making useful comparisons between state standards at one time point. In both the original and the current development, the NAEP scale was used as the benchmark.

The approach developed in this paper arose from the need to deal with practical testing issues. It is well known that, over time, factors such as score inflation, scale drift, differential performance gains, test instrument structure changes, content modification, and style of classroom instruction can all contribute to a deviation in test scores (Thissen, 2007). Apparently, the concept of deviation in test scores is broader than a deviation in proficiency standards.

The entire process of this method involves detecting DPS over time, verifying that standard test conditions have been met, and having the causes of changes evaluated by an expert committee. Under standard conditions, a substantial difference in NAEP equivalents over time indicates possible score inflation. However, the reality of this situation can only be determined after discounting other factors related to changes in test conditions, such as content modification and changes in test instruments.

As mentioned in the section of this paper titled "Evaluating the Test Results," it is possible that differentiation of NAEP equivalents over time may be caused in part by changes in testing conditions and in part by score inflation. The existence of a combination of such causes makes investigation of this matter more difficult. When only limited information is available, individuals wanting to make inferences concerning score inflation with respect to this scenario should do so with due caution.

APPENDIX A: NAEP SAMPLE DESIGN, SCHOOL WEIGHTS, AND TARGET ESTIMATION

NAEP Sample Design and School Weights

State NAEP samples are obtained through a two-stage probability sampling design. The first stage constitutes a probability sample of schools containing the relevant grade. The second stage involves the selection of a random sample of students within each school. To account for the unequal probabilities of selection and to allow for adjustments for non-response, each school and each student is assigned a separate sampling weight.[4] If these weights are not employed in the computation of the statistics of interest, the resulting estimates can be biased.

Because of this caution, appropriate weights are applied in the estimation of the proportion of students in the state above the standard. In general, the school weight equals the inverse of the approximate school selection probability, and the student weight is inversely proportional to the product of the school selection probability and the student selection probability. A more detailed description of school weights can be found in Braun and Qian (2007a).

Because school weights are not retained in the NAEP database, this study computed the school weights in two steps. First, the sum of the student design weights for each school was calculated. This sum was then divided by the number of grade-eligible students.[5]

Details of the creation of school design weights for NAEP can be found in the *NAEP 1998 Technical Report* (Qian, Kaplan, Johnson, Krenzke, & Rust, 2001, Chapter 11).

[4] Students with disabilities and English language learners who cannot be assessed, even with the accommodations that NAEP provides, are not considered non-respondents but are excluded from the population of inference. Their performance is not included in estimates of the NAEP score distributions.

[5] Note that this calculation was carried out only for the subset of NAEP sample schools with complete data. School and student weights were not adjusted for schools lost from the NAEP school sample due to non-response.

The Ratio Estimator for the Target Proportion

Let P_k be the proportion of students achieving the standard at school k, and let w_k be the corresponding school weight. The total number of students meeting the standard is $\sum_{l=1}^{N} P_l \cdot M_l$, where N is the total number of public schools in the state containing the relevant grade and M_l is the number of students who were grade-eligible at school l (including all students with disabilities and English language learners). The statewide target proportion of students meeting the standard is approximately

$$P = \frac{\sum_{l=1}^{N} P_l \cdot M_l}{\sum_{l=1}^{N} M_l} .$$

Horvitz–Thompson estimators (Cochran, 1977) are used to estimate the numerator and denominator of P separately from the state's NAEP school sample. For example, $\sum_{l=1}^{n} w_l M_l$ estimates the total number of eligible students in the state, and $\sum_{l=1}^{n} w_l (P_l \cdot M_l)$ estimates the total number of students meeting the standard. The target proportion, P, of students meeting the standard can be estimated by a ratio estimator:

$$\bar{p}_w = \frac{\sum_{l=1}^{n} w_l (P_l \cdot M_l)}{\sum_{l=1}^{n} w_l M_l} .$$

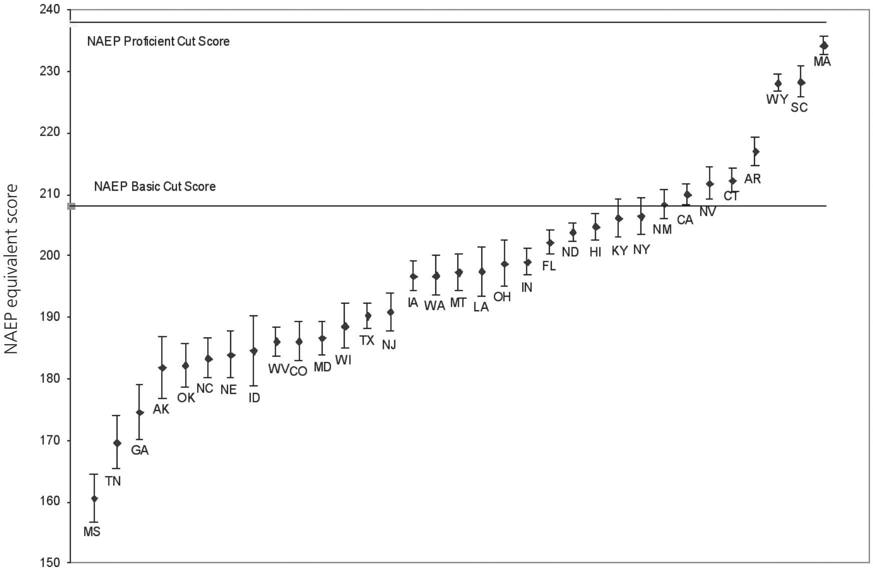
Variance Estimation

When survey variables are observed without error from every respondent in relation to a stratified and clustered sample such as NAEP, the usual complex-sample variance estimators quantify the uncertainty associated with sample statistics (Skinner, Holt, & Smith, 1989). The fact that a specific NAEP score is not assigned to individual students participating in the NAEP assessments (even those who responded to the cognitive items) requires additional statistical analyses to properly quantify the uncertainty associated with inferences about score distributions (Allen et al., 2001; Wolter, 1985).

The total variance of the estimate of the NAEP equivalent to a state standard consists of two components: (a) the error due to sampling schools and students, and (b) the error of measurement that reflects the uncertainty in an assessed student's performance. The sampling error is estimated by applying the jackknife replicate re-sampling (JRR) approach to the mapping procedure. The estimation involves the corresponding schools on the state data and on the NAEP data. The measurement error due to unobservability is estimated by utilizing the variability among the five sets of plausible values generated for each assessed student (Rubin, 1987).

APPENDIX B: RESULTS OF MAPPING STATE STANDARDS FOR THE 2005 NAEP READING AND MATHEMATICS ASSESSMENTS

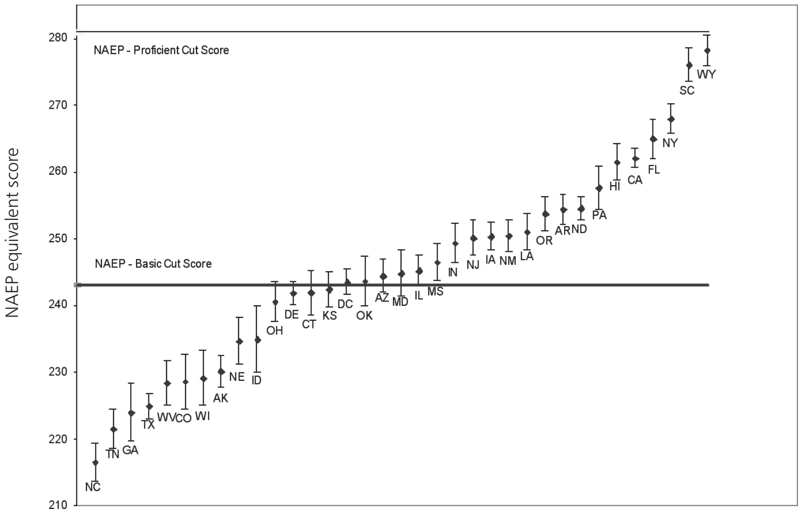
Figure B1: NAEP score equivalents of states' proficiency standards for reading at Grade 4 (2005)



Source:

US Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading Assessment, and National Longitudinal School-Level State Assessment Score Database (NLSLSASD).

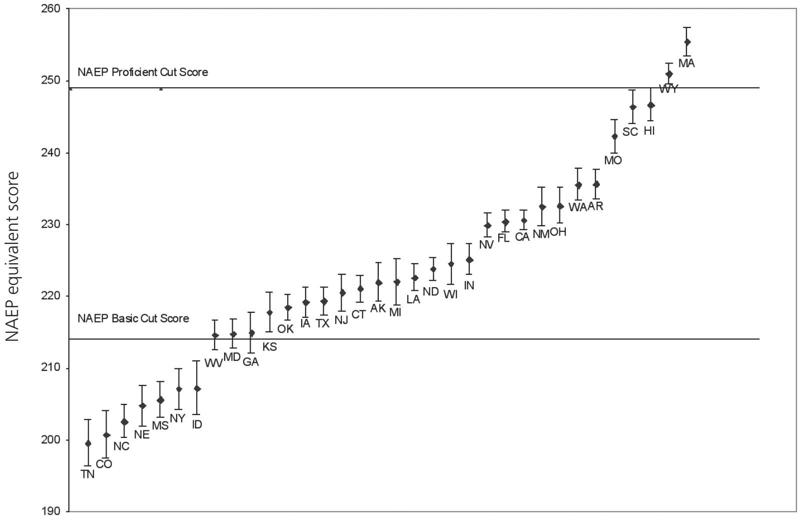
Figure B2: NAEP score equivalents of states' proficiency standards for reading at Grade 8 (2005)



Source:

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading Assessment, and National Longitudinal School-Level State Assessment Score Database (NLSLSASD).

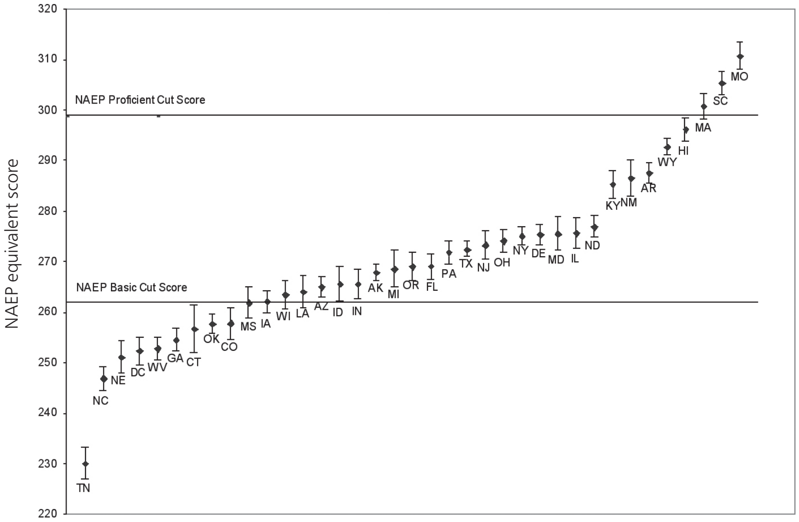
Figure B3: NAEP score equivalents of states' proficiency standards for mathematics at Grade 4 (2005)



Source:

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading Assessment, and National Longitudinal School-Level State Assessment Score Database (NLSLSASD).

Figure B4. NAEP score equivalents of states' proficiency standards for mathematics at Grade 8 (2005)



Source:

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading Assessment, and National Longitudinal School-Level State Assessment Score Database (NLSLSASD).

References

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

American Federation of Teachers. (2006). *Smart testing: Let's get it right.* Unpublished review retrieved from http://www.aft.org/pubs-reports/downloads/teachers/Testingbrief.pdf.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Arenson, K. W. (2004, April 18). Is it grade inflation, or are students just smarter? *New York Times*, p. WK2.

Braun, H. I., & Qian, J. (2007a). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 313–338). New York: Springer.

Braun, H. I., & Qian J. (2007b). *Mapping 2005 state proficiency standards onto the NAEP scales* (NCES Research and Development Report No. NCES 2007–482). Washington DC: National Center for Education Statistics.

Bromley, D. G., Crow, H. L., & Gibson, M. S. (1973). Grade inflation: Trends, causes, and implications. *Phi Delta Kappan, 59*(10), 694–697.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average*. Daniels, WV: Friends for Education.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us* (Rand Corporation Rep. No. MR-924-EDU). Santa Monica, CA: Rand Corporation.

Haberman, S. J. (1978). *Analysis of qualitative data: Vol. 1, Introductory topics*. New York: Academic Press.

Haladyna, T., Nolen, S., & Haas, N. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2−7.

Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., & Phillips, S. E. (1995). *Review of the measurement quality of the Kentucky Instructional Results Information System, 1991−1994*. Frankfort, KY: Office of Education Accountability, Kentucky General Assembly.

Haney, W. (2002). Ensuring failure: How a state's achievement test may be designed to do just that. *Education Week, 56*, 58.

Jones, L., & Olkin, I. (2004). *The nation's report card: Evolution and perspective*s. Bloomington, IN: Phi Delta Kappa International.

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

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives, 8*, 49.

Koretz, D. M. (1988). Arriving in Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator, 12*(2), 8−15, 46−52.

Koretz, D. M. (2005). Alignment, high stakes, and the inflation of test scores. *Yearbook of the National Society for the Study of Education, 104*(2), 99–118.

Koretz, D. M. (2007). Using aggregate-level linkages for estimation and validation: Comments on Thissen and Braun & Qian. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 339–353). New York: Springer.

Koretz, D. M., Bertenthal, M. W., & Green, B. F. (Eds.). (1999). *Embedding questions: The pursuit of a common measure in uncommon tests*. Washington, DC: National Academy of Sciences.

Lewin, T. (2007, June 8). States found to vary widely on education. *New York Times*, p. WK2.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4−16.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that "Everyone is above average." *Educational Measurement: Issues and Practice, 9*(3), 5–14.

Madaus, G. F. (1988a). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education, 65*(3), 29–46.

Madaus, G. F. (1988b). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum* (pp. 83–121). Chicago, IL: University of Chicago Press.

McLaughlin, D., & Bandeira de Mello, V. (2003, June). *Comparing state reading and math performance standards using NAEP*. Paper presented at the National Conference on Large-Scale Assessment, San Antonio, TX.

Neill, M., & the Staff of FairTest. (1997). *Testing our children: A report card on state assessment systems*. Cambridge, MA: National Center for Fair & Open Testing.

Phelps, R. P. (2005). The source of Lake Wobegon. *Third Education Group Review, 1*, 2.

Potter, W. P. (1979). Grade inflation: Unmasking the scourge of the seventies. *College and University, 55*(1), 19–26.

Qian, J., Kaplan, E., Johnson, E., Krenzke, T., & Rust, K. (2001). State weighting procedures and variance estimation. In N. Allen, J. Donoghue, & T. Schoeps (Eds.), *The NAEP 1998 technical report* (pp. 193–225). Washington, DC: National Center for Education Statistics.

Rosovsky, H., & Hartley, M. (2002). *Evaluation and the academy: Are we doing the right thing? Grade inflation and letters of recommendation*. Cambridge, MA: American Academy of Arts and Sciences.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Shepard, L. A. (1988, April). *The harm of measurement-driven instruction.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Skinner, C., Holt, D., & Smith, T. (1989). *Analysis of complex surveys*. New York: John Wiley & Sons.

Smith, M. L. (1991). Meanings of test preparation. *American Educational Research Journal, 28*(3), 521–542.

Thissen, D. (2007). Linking assessments based on aggregate reporting: Background and issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 287–312). New York: Springer.

Wolter, K. (1985). *Introduction to variance estimation.* New York: Springer.

Yates, F. (1934). Contingency table involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society* (Supplement), *1*, 217–235.