Research Report

Mapping State Standards to the NAEP Scale

Henry Braun

Jiahe Qian

November 2008

ETS RR-08-57

Mapping State Standards to the NAEP Scale

Henry Braun[1]
Boston College, Chestnut Hill, MA

Jiahe Qian
ETS, Princeton, NJ

November 2008

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This report describes the derivation and evaluation of a method for comparing the performance standards for public school students set by different states. It is based on an approach proposed by McLaughlin and associates, which constituted an innovative attempt to resolve the confusion and concern that occurs when very different proportions of students in various states are declared to have met a standard with the same label. Our method, like McLaughlin's, employs equipercentile methods to map state standards on to a common scale, that associated with the National Assessment of Educational Progress (NAEP). We have also derived error estimates that take into account both NAEP's complex sampling design and measurement errors. The method was applied to two data sets, and the results were qualitatively similar to those obtained by McLaughlin's method. The paper notes the superior statistical properties of the proposed method and presents evidence that supports the viability and general utility of this approach.

Key words: NAEP, proficiency standards, state tests, equipercentile linking, No Child Left Behind

Acknowledgments

The authors thank Mary Pitoniak and Bruce Kaplan for help in planning this study, as well as John Wiley and Sailesh Vezzu for assistance with computing. They also appreciate the suggestions and comments by James Carlson, Andrew Kolstad, Taslima Rahman, Alexandra Sedlacek, Bruce Kaplan, Paul Holland, John Mazzeo, Don McLaughlin, and the editors. They are particularly grateful to Dan Koretz for his unflagging efforts to clarify the import of our findings. Of course, any remaining errors or misinterpretations are the responsibility of the authors. This work was supported by the National Center for Education Statistics, Contract # ED-02-CO-0023.

Table of Contents

List of Tables

Page

List of Figures

1. Introduction

During the 1990s, under the impetus of standards-based reform, many states established performance standards for their students in selected grades and subjects. Under the most recent reauthorization of the Elementary and Secondary Education Act (ESEA), the No Child Left Behind Act (NCLB), all states are required to set such standards in reading and mathematics for Grades 3-8 and also for at least one grade in high school. NCLB, however, leaves to the states the responsibility of determining the curriculum, selecting the assessments and setting challenging academic standards. Not surprisingly, the result has been substantial heterogeneity in both the quality and apparent stringency of the standards set by the states (Lane, 2004; Linn, 2003). One consequence is that, in a particular grade, very different proportions of students in the various states have been declared to have met a standard with the same label (e.g., *proficient*). These differences have occasioned much confusion and concern among stakeholders.

A moment's reflection shows that unambiguous comparisons of standards among states are problematic in view of the flexibility accorded to the states under NCLB. That is, if states were using the same test, then determining the relative stringency of the standards could be accomplished by simply comparing the cut-points established by each state. In the present context, such direct comparisons are impossible. It is evident that there would be value in somehow placing all state standards on a common basis to facilitate approximate but credible comparisons in student test performance. As will be made clear below, any such effort cannot eliminate an essential indeterminacy that must be taken into account in the interpretation of the results. Nonetheless, given both the importance and the visibility of the issue, it seems appropriate to address it as responsibly as one can.

In the past, there have been a number of calls to somehow link all the states' test score scales directly or, failing that, to map them all on to the National Assessment of Educational Progress (NAEP) scale, inasmuch as NAEP is the only test that is administered in a uniform manner across states. (Moreover, NAEP is generally regarded as meeting high standards with respect to test design, test content, and psychometric quality.[2]) If such linkages were possible, then comparisons among state standards would be relatively straightforward. Unfortunately, the literature is replete with arguments against the appropriateness of such mappings (e.g., Linn, 1993). More recently, two studies carried out under the auspices of the National Research Council (NRC; Feuer, Holland, Green, Bertenthal, & Hemphill, 1998; Koretz, Bertenthal, & Green, 1999)

concluded that, for a number of reasons, mappings could not be validly constructed at the student level.

McLaughlin and his associates (McLaughlin & Bandeira de Mello, 2002, 2003) made an innovative attempt to circumvent some of the difficulties cited in the NRC studies. Their approach was to carry out the mapping to the NAEP scale only at the school level (at a single point) and then, by aggregation, to the state level. Specifically, they employed equipercentile linking (Braun & Holland, 1982) in each school to find a point on the NAEP score scale that best corresponds to the state standard. That point represents the local estimate of the state standard on the NAEP scale. A simple average of these local estimates across all the schools in the NAEP sample (approximately 80-100 schools) yields the final estimate of the NAEP equivalent to the state standard. It should be noted that in their computations they used the so-called full population estimates (FPE) of NAEP score distributions (McLaughlin, 2000; Pitoniak & Mead, 2003), rather than the reported NAEP distributions. Evidence for the plausibility of most of these mappings of the state standards to the NAEP scale can be found in McLaughlin and Bandeira de Mello (2002).

This paper presents an alternative approach, albeit one that also relies on equipercentile linking. This method takes into account NAEP's complex sample design, both in obtaining an estimate of the NAEP equivalent of a state standard and in deriving an estimated variance of the NAEP equivalent. The method was applied to data from states' 2000 mathematics assessment and the NAEP 2000 mathematics assessment, as well as to data from states' 2002 reading assessment and the NAEP 2002 reading assessment.[3] Aside from our use of the reported NAEP distributions, the main difference between the approach adopted in this study and that of McLaughlin and his associates is that we obtained what might be termed a direct estimate of the NAEP equivalent by using appropriately weighted estimates of the state's NAEP distribution and of the proportion of students meeting the state's achievement standard(s). The rationale is that such a direct estimate should be more precise than one that relies on a simple average of a large number of less precise estimates from a probability sample of schools.

This paper reports the results of a number of data analyses that support, on methodological grounds, a preference for this approach to that of McLaughlin and Bandeira de Mello (2003). We then assert that most of the observed differences among states in the proportions of students meeting states' proficiency standards are the result of differences in the stringency of their standards. This is followed by an examination of the evidence for the assertion. If it is essentially

correct, then it has important implications for education policy. In particular, it begs the question of whether all students deemed proficient are actually prepared to succeed once they leave the public school system.

Underlying both the approach described here and McLaughlin's approach is the assumption that, for a particular subject and grade, the state tests and NAEP are similar in content and structure. This is necessary so that the linking is not simply a meaningless exercise in numerology. It is also worth noting that both approaches treat as equivalent the proportions meeting a standard defined in terms of an estimate of the state score distribution and a cut point defined in terms of an estimate of a NAEP score distribution. These two estimates are based on data at different levels of analysis—the former on cumulating scores of individual students and the latter on obtaining a direct estimate of the underlying true score distribution. Of course, there are also differences in the use of a census rather than a sample, in exclusion rules, in the kinds of instruments used, and so on. However, inasmuch as state test forms are usually fairly long and have reasonably high reliabilities, we believe that for our purposes we can ignore these differences.

It should be emphasized that the location of the NAEP score equivalent of a state's proficiency standard is not simply a function of the placement of the state's standard on the state's own test score scale. Rather, it also depends on the curriculum delivered to students across the state and the test's coverage of that curriculum with respect to both breadth and depth, as well as the relationship of both to the NAEP framework and the NAEP assessment administered to students. Thus, the variation among states' NAEP equivalent scores reflects the interaction of multiple factors, which can complicate interpretation of the results.

In the next two sections we will provide a brief outline and a more detailed description of the proposed method. Section 4 describes the derivation of the variance estimates, and section 5 presents results for Grade 4 in both mathematics and reading. Specifically, state-by-state results are presented for standards labeled proficient or its equivalent. Section 6 describes mapping NAEP standards into a state scale, and the final section offers conclusions and recommendations.

2. Outline of the ETS Procedure for Mapping State Standards to NAEP Scale

The procedure is carried out separately for each state that participated in NAEP and is represented in the National Longitudinal School-level State Assessment Score Database (for the corresponding academic year) referenced above. In our description of the procedure, we will refer to the mathematics data. An identical procedure was used for the reading data. Let *P*, which is

formally defined in Section 3.2, denote the state-wide proportion of students meeting a particular standard. To emphasize the differences in the two approaches, we will refer to our method as *weighted aggregate mapping* (WAM) and that of McLaughlin and associates as *unweighted local mapping* (ULM).

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, we identified the schools in the state NAEP sample and matched them with their records in the National Longitudinal School-Level State Assessment Score Database.[4] For each school, we obtained the proportion of students meeting the state standard. Using the school weights from the NAEP design, we obtained an estimate of *P* using a ratio estimator, $\overline{p}_w$, 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. In Section 3.1, we will describe the weights and the ratio estimator in more detail.

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 is the procedure that was carried out to generate the results contained in the report that follows each NAEP assessment. Let $\hat{F}$ denote the estimated distribution.

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 the proportion *P* of students meeting the state's own performance standard (defined with respect to the state test score scale) was estimated by $\overline{p}_w$ and the NAEP score distribution was calculated, as in Steps 1 and 2 above, we mapped the performance standard to the NAEP scale, by finding the point y_{WAM} on the NAEP scale that is the $(1-\overline{p}_w)$th quantile:

$$y_{\text{WAM}} = \hat{F}^{-1}\left(1-\overline{p}_w\right).$$

We took y_{WAM} to be the estimated NAEP equivalent score to the state standard. If the state employs more than one standard, the above procedure can be repeated for each one.

4. *Compute the variance of the estimated NAEP equivalent*.
   Using the jackknife procedure, we estimated the contribution of the sampling of schools and students to the variance of the estimator and combined it with an estimate of the contribution of measurement error to obtain a total variance estimate.

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 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, only $\hat{p}_w$ need be obtained 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 based on students in the state's NAEP sample who took the NAEP assessment. The estimated NAEP equivalent to the state standard, y_{WAM}, is the point on the NAEP scale, such that the corresponding upper tail area of the NAEP distribution also equals $\hat{p}_w$. It should be clear that, for a given distribution of state test scores and a particular distribution of NAEP scores, a larger $\hat{p}_w$ corresponds to a lower y_{WAM}, and vice-versa.

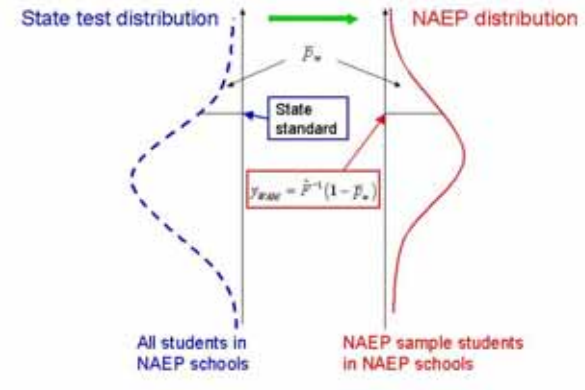

Figure 1. **The schematic of the mapping procedure.**

The approach of McLaughlin and associates (2002 & 2003) can be described in a series of steps analogous to those of the ETS procedure:

1. *For each school in the NAEP sample, obtain (from the NLSLSASD) the proportion of students in that school who meet the state performance standard on the state's own assessment.*

   This proportion is denoted by P_k for school *k*. It is based on the scores of all students in the school who took the state test (typically, nearly all the students in the relevant grade).

2. *For each school in the NAEP sample, estimate the distribution of scores on the NAEP assessment for that school.*

   This distribution is based on scores on the NAEP assessment in the school, including imputing scores for those who were sampled but excluded from the assessment. This estimate is referred to as a full population estimate (FPE). This estimate is based on the scores of those students in the school who took the NAEP assessment (typically, 20 students in the relevant grade).

3. *For each school in the NAEP sample, find the point on the NAEP score scale at which the estimated proportion of students in that school scoring above that point equals the proportion of students in that school meeting the state's own performance standard.*

   This point is the NAEP equivalent for the school and is denoted by z_k. It is the $(1-P_k)$th quantile, $\hat{F}^{-1}(1-P_k)$, of the FPE of the NAEP distribution in the school.

4. *Average the NAEP equivalents for the schools in the NAEP sample to estimate the NAEP equivalent for the state.*

   That is, compute

$$\bar{z}_{ULM} = \frac{1}{n}\sum_{k=1}^{n} z_k = \frac{1}{n}\sum_{k=1}^{n} \hat{F}^{-1}(1-P_k).$$

The ETS procedure estimates the NAEP equivalent to the state standard, y_{WAM}, by using a single aggregate distribution of all student scores in all NAEP selected schools in a state. The procedure introduced by McLaughlin and associates use the average of mapped standards of approximately 100 single-school, $\bar{z}_{ULM}$, to estimate the target statistic. While the empirical results generally show small differences between y_{WAM} and $\bar{z}_{ULM}$, there are important conceptual differences. For a detailed explanation, see the appendix.

3. Details of the Methodology

3.1 The Weights for NAEP Schools

The state NAEP sample was obtained through a two-stage probability sampling design. The first stage constituted a probability sample of schools containing the relevant grade. The second stage involved the selection of a random sample of students within schools.

To account for the unequal probabilities of selection, and to allow for adjustments for nonresponse, each school and each student were assigned separate sampling weights. If these weights are not applied in the computation of the statistics of interest, the resulting estimates can be biased. With this caution in mind, we applied appropriate weights in the estimation of the proportion of students in the state above the standard. In general, the student weight is inversely proportional to the product of the school selection probability and the student selection probability.

Formally, let *N* be the total number of schools in a state and M_k be the number of students who were grade-eligible at school *k.* Therefore, the total number of eligible students in the state is $\sum_{l=1}^{N} M_l$. Let *n* be the number of schools in the state NAEP sample. Let π_k be the school selection probability, which is proportional to its size M_k, and let $\pi_{i|k}$ be the conditional probability of selection for student *i* in school *k*. Suppose that *b* students are randomly selected from school *k*. Then the unconditional selection probability of student *i* in school *k* is

$$\pi_{ki} = \pi_k \cdot \pi_{i|k} = \frac{a \cdot M_k}{\sum_{l=1}^{N} M_l} \cdot \frac{b}{M_k},$$

where *a* is a constant of normalization. Then the weight of student *i* in school *k* is

$$w_{ki} = w_k \cdot w_{i|k} = \frac{1}{a \cdot M_k / \sum_{l=1}^{N} M_l} \cdot \frac{1}{b / M_k}.$$

This formula is only an approximation, since students are selected without replacement and the vicissitudes of field work necessitate modifications to the ideal weights. For example, nonresponse adjustments to the weights are employed in NAEP to account for effects of schools and students who were selected but did not participate. In any case, the weight of school *k* in a state NAEP sample is approximately

$$w_k = \frac{1}{a \cdot M_k / \sum_{l=1}^{N} M_l},$$

which equals the inverse of the approximate school selection probability. Since school weights are not retained in the NAEP database, for this study the estimates of school weights were computed in two steps. First the sum of the student design weights for each school was calculated, and then this sum was divided by the number of eligible students. Details of the creation of school design weights for NAEP can be found in NAEP 1998 Technical Report, Chapter 11 (Qian, Kaplan, Johnson, Krenzke, & Rust, 2001).

3.2 The Ratio Estimator for the Target Proportion

Let P_k be the proportion of students achieving the standard at school *k*. The total number of students meeting the standard is $\sum_{l=1}^{N} P_l \cdot M_l$. The statewide target proportion of students meeting the standard is approximately

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

Using Horvitz–Thompson estimators, the numerator and denominator of *P* are estimated 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 $\sum_{l=1}^{n} w_l \left(P_l \cdot M_l \right)$ 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:

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

The Horvitz-Thompson estimators, $\sum_{l=1}^{n} w_l M_l$ and $\sum_{l=1}^{n} w_l \left(P_l \cdot M_l \right)$, are unbiased estimators of the corresponding population totals. Nevertheless, the ratio estimator $\overline{p}_w$ is biased with an order of $O\left(1/n\right)$ (Cochran, 1977).

An interesting result can be derived by substituting for the school weight w_l in $\overline{p}_w$ the inverse of the school selection probability. Simple algebra shows that the corresponding estimate reduces to $(1/n)\sum_{l=1}^{n} P_l$, which is denoted by $\overline{p}$. Thus, with this simplification, the ratio estimator equals the simple average of P_k in the sample. Because the weights in NAEP samples reflect the effects of oversampling, nonresponse adjustments, and trimming, the actual school weight, w_k, will differ somewhat from $\sum_{l=1}^{N} M_l /(a \cdot M_l)$, and therefore, $\overline{p}_w$ will also differ slightly from $\overline{p}$. However, since the school size data are not available for all schools in the states in the study, we have chosen to replace *P* by $\overline{P}$, the population analog of $\overline{p}$; that is $\overline{P} = (1/N)\sum_{l=1}^{N} P_l$.

We have chosen to use the ratio estimator $\overline{p}_w$ in our analysis. A plausible alternative would be *P*, which is based on data from all the schools in the state containing the relevant grade. With our choice, the same schools contribute to the estimation of the relevant parameters of the state test score distribution and the NAEP score distribution. We believe that this match is more consistent with the logic underlying McLaughlin's method and should yield results with smaller mean squared error. As we see below, the differences between $\overline{p}_w$ and $\overline{P}$ are typically very small.[5]

It is worth noting that both the approach suggested here and the one developed by McLaughlin treat as equivalent the proportions meeting a standard defined in terms of an estimate of the state score distribution and a cut-point defined in terms of an estimate of a NAEP score distribution. These two estimates are based on different principles: The former on cumulating (estimated) scores of individual students and the latter on obtaining a direct estimate of the underlying true score distribution. Of course, there are also differences in the use of a census rather than a sample, in exclusion rules, in the kinds of instruments used, and so on. However, inasmuch as state test forms are usually fairly long and have reasonably high reliabilities, we believe that for our purposes we can ignore these differences.

3.3 Empirical Evaluation of the Estimates

3.3.1 Data Resources Used in Analysis

The data analyzed in this study consisted of two sets of NAEP data: (a) the NAEP 2000 mathematics assessments for Grade 4 and Grade 8 students in the R3 [6] sample, and (b) the NAEP 2002 reading assessments for Grade 4 and Grade 8 students in the R3 sample. We also employed two sets of state test data: (a) 2000 state mathematics tests and (b) 2002 state reading tests. The state data were obtained from the NLSLSASD database. 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.

3.3.2 Evaluation of the Bias of the Estimates of the Target Proportions

The state standards for the 2000 state mathematics tests. We evaluated the approximate bias of the sample estimates of the proportion proficient by analyzing the Grade 4 (G4) 2000 mathematics standards. We compared the ratio estimator, $\overline{p}_w$, and the ordinary simple average of school proportions, $\overline{p}$, to the statewide target proportion of students meeting the standard, $\overline{P}$, which was defined in the previous section. For present purposes, $\overline{P}$ is treated as the true state percentage.

Tables 1 and 2 summarize, for each state standard, the key statistics of the 2000 state mathematics test score distribution. The first and second columns of the tables contain the total number of (grade-relevant) schools in the state population and the number of NAEP schools in the

sample for each state. The third column is the state-wide target proportion of students meeting the standard. The fourth and fifth columns are the estimates denoted by $\overline{p}_w$ and $\overline{p}$.

We defined the bias of the estimators $\overline{p}_w$ and $\overline{p}$ as ($\overline{p}_w$ - $\overline{P}$) and ($\overline{p}$ - $\overline{P}$), respectively. The biases of both estimators are small: For G4 of 2000 mathematics, the bias of $\overline{p}$ is larger than the bias of $\overline{p}_w$ for 28 out of 46 state standards. The averages of the absolute biases for $\overline{p}$ and $\overline{p}_w$ are 1.1% and 1.2%, and the maxima of the absolute biases are 6.7% and 6.0%, respectively. The average of the differences between two estimators is just 0.9%. For G8 of 2000 mathematics, the bias of $\overline{p}$ is larger than the bias of $\overline{p}_w$ for 23 out of 53 state standards. The average of the absolute biases for $\overline{p}$ and $\overline{p}_w$ are 0.9% and 0.7%. The maxima of the absolute biases are 4.5% and 3.4% for $\overline{p}$ and $\overline{p}_w$ respectively. The average of the differences between two estimators is 0.7%. Thus both $\overline{p}_w$ and $\overline{p}$ are only slightly biased estimates of $\overline{P}$.

The state standards for the 2002 state reading tests. This section presents the results for the G4 and G8 2002 reading standards. Tables 3 and 4, with the same format as Tables 1 and 2, summarize the 5 key statistics of the 2002 state reading test score distributions. As was the case for mathematics, the biases of both estimators, $\overline{p}$ and $\overline{p}_w$, are also small. For G4, the bias of $\overline{p}$ is larger than the bias of $\overline{p}_w$ for 24 out of 51 state standards. The averages of the absolute bias for $\overline{p}$ and $\overline{p}_w$ are 1.4% and 1.1%, and the maxima of the absolute biases are 9.0% and 3.7%, respectively. The average of the differences between two estimators is about 1.2%. For G8, the bias of $\overline{p}$ is larger than the bias of $\overline{p}_w$ for 25 out of 59 state standards. The averages of the absolute bias for $\overline{p}$ and $\overline{p}_w$ are 1.9% and 1.8%. The maxima of the absolute biases are 11.9% and 12.5% for $\overline{p}$ and $\overline{p}_w$ respectively. The average of the differences between two estimators is 1.2%.

3.3.3 Evaluation of the Estimates of the NAEP Equivalent to the State Standard

Because the target quantity (the NAEP equivalent to the state standard) is not known, it is difficult to determine the bias of any estimate. However, both sampling theory and general NAEP empirical results indicate that estimates using design weights provide superior results to those that don't. The estimate y_{WAM} defined in Section 2 does employ these design weights.

Table 1

G4 2000 Math: The Unweighted and Weighted Proportions of Tested Students With Scores at or Above the State Standards

State & standard	Total # of schools in state (1)	# of schools in NAEP sample (2)	State school population: Proportion of students meeting the standard, $\overline{P}$ (3)	NAEP school sample: Weighted average proportion meeting the standard, $\overline{p}_w$ (4)	NAEP school sample: Unweighted average proportion meeting the standard, $\overline{p}$ (5)
AR benchmark	511	94	.37	.36	.35
CA PR25	4,827	77	.72	.74	.73
CA PR50	4,827	77	.50	.53	.51
CA PR75	4,827	77	.28	.31	.30
CT goal	589	105	.64	.64	.62
GA meets	999	98	.62	.62	.61
GA exceeds	999	98	.11	.10	.10
KS basic	741	75	.85	.84	.84
KS satisfactory	741	75	.60	.59	.58
KS proficient	741	75	.37	.36	.34
KS advanced	741	75	.13	.12	.11
LA appro. basic	787	106	.70	.74	.73
LA basic	787	106	.47	.50	.49
LA proficient	787	106	.11	.11	.11
LA advanced	787	106	.01	.01	.01
MA pass	1,020	105	.82	.83	.81
MA proficient	1,020	105	.40	.42	.40
MA advanced	1,020	105	.11	.12	.11
ME partially meets	343	105	.71	.72	.71
ME meets	343	105	.23	.24	.23
ME exceeds	343	105	.02	.02	.01
MI moderate	1,910	84	.91	.93	.93
MI satisfactory	1,910	84	.75	.77	.76

(Table continues)

Table 1 (continued)

State & standard	Total # of schools in state (1)	# of schools in NAEP sample (2)	State school population: Proportion of students meeting the standard, $\overline{P}$ (3)	NAEP school sample: Weighted average proportion meeting the standard, $\overline{p}_w$ (4)	NAEP school sample: Unweighted average proportion meeting the standard, $\overline{p}$ (5)
MO progressing	1,097	99	.97	.97	.97
MO near proficient	1,097	99	.78	.79	.78
MO proficient	1,097	99	.37	.36	.36
MO advanced	1,097	99	.08	.08	.07
NC inconsist. mastery	1,229	107	.98	.98	.98
NC consist. mastery	1,229	107	.84	.85	.84
NC superior	1,229	107	.41	.40	.40
NE emerging	161	17	.88	.87	.88
NE proficient	161	17	.64	.60	.62
NE advanced	161	17	.37	.32	.34
NY needs improvement	1,476	40	.92	.93	.93
NY meets	1,476	40	.67	.68	.67
NY exceeds	1,476	40	.20	.20	.19
OH pass	1,990	84	.49	.43	.42
RI proficient	188	108	.20	.21	.20
SC basic	549	101	.62	.62	.62
SC proficient	549	101	.24	.23	.23
SC advanced	549	101	.08	.07	.07
TX pass	3,417	99	.87	.89	.88
VT meets	213	60	.69	.69	.68
WY partial proficient	162	83	.62	.61	.61
WY proficient	162	83	.27	.26	.27
WY advanced	162	83	.05	.05	.05

Table 2

G8 2000 Math: The Unweighted and Weighted Proportions of Tested Students With Scores at or Above the State Standards

State & standard	Total # of schools in state (1)	# of schools in NAEP sample (2)	State school population: Proportion of students meeting the standard, $\overline{P}$ (3)	NAEP school sample: Weighted average proportion meeting the standard, $\overline{p}_w$ (4)	NAEP school sample: Unweighted average proportion meeting the standard, $\overline{p}$ (5)
AZ approach	408	76	.51	.54	.54
AZ meets	408	76	.16	.18	.18
AZ exceeds	408	76	.05	.06	.06
CA PR25	1748	75	.70	.71	.70
CA PR50	1748	75	.48	.48	.46
CA PR75	1748	75	.23	.23	.22
CT at goal	227	102	.58	.59	.59
GA meets	368	97	54	.55	.54
GA exceeds	368	97	.11	.11	.11
HI Stanine 5	50	46	.61	.60	.58
IL meets	1,368	78	.46	.47	.41
IL exceeds	1,368	78	.11	.11	.09
IN meets	449	76	.63	.65	.65
LA approach basic	460	102	.67	.70	.70
LA basic	460	102	.46	.48	.48
LA proficient	460	102	.07	.07	.07
LA advanced	460	102	.03	.03	.03
MA pass	402	96	.61	.62	.62
MA proficient	402	96	.34	.35	.35
MA advanced	402	96	.10	.10	.11
MD satisfactory	264	102	.50	.51	.50
MD excellent	264	102	.15	.16	.15
ME partial	215	80	.60	.60	.60
ME meets	215	80	.21	.21	.20
ME exceeds	215	80	.01	.01	.01
MN pass	459	59	.72	.73	.73

(Table continues)

Table 2 (continued)

State & standard	Total # of schools in state (1)	# of schools in NAEP sample (2)	State school population: Proportion of students meeting the standard, $\overline{P}$ (3)	NAEP school sample: Weighted average proportion meeting the standard, $\overline{p}_w$ (4)	NAEP school sample: Unweighted average proportion meeting the standard, $\overline{p}$ (5)
MO progressing	623	99	.78	.76	.74
MO near proficient	623	99	.43	.41	.40
MO proficient	623	99	.14	.13	.13
MO advanced	623	99	.01	.01	.01
NC inconsist. mastery	600	103	.95	.96	.95
NC consist mastery	600	103	.81	.81	.80
NC superior	600	13	.44	.44	.44
NV PR25	74	52	.74	.74	.74
NV PR75	74	52	.24	.25	.24
NY needs improve.	693	45	.77	.79	.78
NY meets	693	45	.42	.42	.41
NY exceeds	693	45	.07	.06	.06
OK little knowledge	551	109	.88	.88	.88
OK satisfactory	551	109	.71	.71	.70
OK advanced	551	109	.13	.13	.12
OR meets	297	77	.56	.56	.55
OR exceeds	297	77	.30	.30	.30
RI meets	53	51	.27	.27	.28
SC basic	251	93	.63	.63	.63
SC proficient	251	93	.20	.19	.19
SC advanced	251	93	.07	.06	.06
TX pass	1,571	103	.90	.90	.90
VA pass	367	104	.61	.62	.61
VT meets	124	71	.67	.66	.65
WY partial	79	60	.70	.70	.69
WY proficient	79	60	.32	.32	.31
WY advanced	79	60	.09	.09	.08

Table 3

G4 2002 Reading: The Weighted Proportions of Tested Students With Scores at or Above the State Standards

State & standard	Total # of schools in state (1)	# of schools in NAEP sample (2)	State school population: Proportion of students meeting the standard, $\bar{P}$ (3)	NAEP school sample: Weighted average proportion meeting the standard, $\bar{p}_w$ (4)	NAEP school sample: Unweighted average proportion meeting the standard, $\bar{p}$ (5)
AR proficient	516	105	.56	.56	.55
AR advanced	516	105	.05	.05	.04
CA PR25	5,089	140	.75	.73	.68
CA PR50	5,089	140	.51	.49	.42
CA PR75	5,089	140	.28	.26	.21
CT Level 1	585	108	.79	.82	.81
CT Level 2	585	108	.69	.73	.71
CT Level 3	585	108	.56	.59	.58
FL Level 1	1,694	103	.69	.69	.68
FL Level 2	1,694	103	.53	.53	.53
FL Level 3	1,694	103	.26	.26	.26
FL Level 4	1,694	103	.06	.06	.06
GA meets standard	1,064	137	.79	.79	.76
GA exceeds standard	1,064	137	.37	.40	.36
ME partially meets standard	354	96	.90	.92	.91
ME meets standard	354	96	.46	.50	.48
ME exceeds standard	354	96	.01	.01	.01
MA pct passing	1,035	110	.90	.92	.91
MA pct proficient	1,035	110	.52	.55	.54
MA pct advanced	1,035	110	.07	.08	.07
MI pct moderate	1,963	108	.80	.81	.79
MI pct satisfactory	1,963	108	.56	.57	.55
MS pct basic	450	98	.90	.91	.91
MS pct proficient	450	98	.83	.83	.83

(Table continues)

Table 3 (continued)

State & standard	Total # of schools in state (1)	# of schools in NAEP sample (2)	State school population: Proportion of students meeting the standard, $\bar{P}$ (3)	NAEP school sample: Weighted average proportion meeting the standard, $\bar{p}_w$ (4)	NAEP school sample: Unweighted average proportion meeting the standard, $\bar{p}$ (5)
MT pct at + near proficient	284	65	.90	.90	.92
MT pct at + proficient	284	65	.77	.78	.80
MT pct advanced	284	65	.19	.18	.19
NY Level 1	2,289	88	.92	.92	.91
NY Level 2	2,289	88	.63	.62	.60
NY Level 3	2,289	88	.21	.21	.19
NC Level 1	1,262	111	.95	.95	.95
NC Level 2	1,262	111	.76	.76	.76
NC Level 3	1,262	111	.31	.32	.31
OH pct passing	1,991	106	.67	.68	.67
RI pct prof. (analysis)	190	110	.62	.61	.60
RI pct prof. (basic)	190	110	.75	.73	.73
SC pct passing	569	104	.79	.79	.79
SC pct proficient	569	104	.32	.33	.32
SC pct advanced	569	104	.02	.02	.02
TX pct passing	3,598	136	.92	.92	.92
TX pct mastering	3,598	136	.48	.50	.47
VT pct meet basic	215	98	.80	.80	.79
WA Level 1	1,074	83	.94	.96	.96
WA Level 2	1,074	83	.66	.69	.69
WA Level 3	1,074	83	.27	.28	.28
WI pct basic	1,128	61	.94	.94	.94
WI pct proficient	1,128	61	.83	.83	.82
WI pct advanced	1,128	61	.19	.19	.18
WY pct partial proficient	156	143	.80	.80	.80
WY pct above proficient	156	143	.44	.44	.43
WY pct advanced	156	143	.14	.13	.13

Table 4

G8 2002 Reading: The Weighted Proportions of Tested Students With Scores at or Above the State Standards

State & standard	Total # of schools in state (1)	# of schools in NAEP sample (2)	State school population: Proportion of students meeting the standard, $\overline{P}$ (3)	NAEP school sample: Weighted average proportion meeting the standard, $\overline{p}_w$ (4)	NAEP school sample: Unweighted average proportion meeting the standard, $\overline{p}$ (5)
AR basic	337	101	.84	.85	.85
AR proficient	337	101	.30	.31	.31
AR advanced	337	101	.03	.03	.03
CA PR25	1,943	122	.74	.73	.66
CA PR50	1,943	122	.50	.48	.41
CA PR75	1,943	122	.21	.20	.16
CT Level 1	227	103	.84	.84	.85
CT Level 2	227	103	.76	.77	.77
CT Level 3	227	103	.65	.66	.67
DE below standard	35	31	.90	.89	.90
DE meets standard	35	31	.73	.71	.74
DE exceeds standard	35	31	.11	.11	.12
DE distinguished	35	31	.04	.03	.04
FL Level 1	692	102	.66	.74	.73
FL Level 2	692	102	.41	.48	.47
FL Level 3	692	102	.15	.18	.18
FL Level 4	692	102	.03	.03	.03
GA meets standard	439	105	.79	.81	.80
GA exceeds standard	439	105	.41	.45	.42
HI Stanine 4+	58	50	.73	.75	.73
HI Stanine 5+	58	50	.52	.54	.52
Hi Stanine 7+	58	50	.21	.22	.21
IL meets standard	1,404	106	.66	.69	.60
IL exceeds standard	1,404	106	.08	.11	.08

(Table continues)

Table 4 (continued)

State & standard	Total # of schools in state (1)	# of schools in NAEP sample (2)	State school population: Proportion of students meeting the standard, $\overline{P}$ (3)	NAEP school sample: Weighted average proportion meeting the standard, $\overline{p}_w$ (4)	NAEP school sample: Unweighted average proportion meeting the standard, $\overline{p}$ (5)
IN pct at or above	448	100	.67	.68	.67
KS pct basic	400	80	.89	.89	.89
KS pct satisfactory	400	80	.66	.66	.65
KS pct proficient	400	80	.37	.37	.36
KS pct advanced	400	80	.08	.08	.08
ME partially meets standard	220	98	.87	.87	.87
ME meets standard	220	98	.42	.44	.43
ME exceeds standard	200	98	.01	.02	.01
MD pct satisfactory	151	54	.22	.23	.23
MD pct excellent	151	54	.03	.03	.03
MS pct basic	294	93	.74	.76	.76
MS pct proficient	294	93	.46	.49	.49
MT pct at + near proficient	211	69	.86	.87	.86
MT pct at + proficient	211	69	.72	.73	.72
MT pct advanced	211	69	.15	.16	.16
NY Level 1	1,129	80	.92	.94	.93
NY Level 2	1,129	80	.43	.43	.42
NY Level 3	1,129	80	.10	.09	.09
NC Level 1	622	106	.97	.98	.98
NC Level 2	622	106	.81	.85	.85
NC Level 3	622	106	.36	.41	.40
OR pct meets or exceeds	315	82	.62	.66	.65
OR pct exceeds	315	82	.31	.35	.34

(Table continues)

Table 4 (continued)

State & standard	Total # of schools in state (1)	# of schools in NAEP sample (2)	State school population: Proportion of students meeting the standard, $\overline{P}$ (3)	NAEP school sample: Weighted average proportion meeting the standard, $\overline{p}_w$ (4)	NAEP school sample: Unweighted average proportion meeting the standard, $\overline{p}$ (5)
PA pct basic	800	101	.77	.80	.80
PA pct proficient	800	101	.55	.59	.58
PA pct advanced	800	101	.18	.20	.20
RI pct prof. (analysis)	55	53	.27	.27	.28
SC pct passing	263	95	.68	.69	.68
SC pct proficient	263	95	.25	.26	.25
SC pct advanced	263	95	.04	.04	.04
TX pct passing	1,662	123	.94	.95	.94
TX pct mastering	1,662	123	.57	.57	.53
VT pct meets basic	126	96	.65	.64	.64
VA pct passing	441	101	.59	.71	.71
WY pct partial proficient	78	71	.80	.79	.80
WY pct above proficient	78	71	.41	.38	.40
WY pct advanced	78	71	.07	.07	.07

In McLaughlin's analysis, full population estimates (FPE) of the NAEP scale score distribution were used to adjust the estimated NAEP score distribution to account for the exclusion of some students with disabilities (SD) or limited English proficiency (LEP) in the NAEP assessments. The FPE approach requires the imputation of the performance of those excluded students (McLaughlin, 2000). Since the imputed scale scores for excluded SD/LEP students usually fall at the low end of the distribution, the FPE of the NAEP distribution is stochastically smaller than the reported NAEP distribution. To study the effect of employing the FPE, we also applied McLaughin's procedure to the reported NAEP distribution. The symbol $\overline{z}'_{ULM}$ denotes the results employing the FPE adjustment of the NAEP distributions. Note that the symbol $\overline{z}_{ULM}$,

defined in Section 2, denotes the ULM estimates based on the reported NAEP distributions. Our comparison focused on the results based on the 2000 mathematics data.

Tables 5 and 6 present three estimates of the NAEP equivalents to the state standards for Grade 4 and Grade 8 of the 2000 state mathematics tests. Columns 1 and 2 contain the results for $\bar{z}_{ULM}$ and $\bar{z}'_{ULM}$, while column 3 presents those for y_{WAM}.

For G4 of 2000 mathematics, on average, y_{WAM} is about 0.5 points higher than $\bar{z}_{ULM}$, but about 1.5 points lower than the mean of $\bar{z}'_{ULM}$, which is 228.4. Apparently, in this setting, the use of design weights has an effect similar to the use of the FPE. Overall, of 46 state standards, for about 61%, $\bar{z}_{ULM}$ is lower than y_{WAM}, and for about 72%, y_{WAM} is lower than $\bar{z}'_{ULM}$.

For G8 of 2000 mathematics, on average, y_{WAM} is 2.1 points higher than $\bar{z}_{ULM}$ but 0.6 points lower than the mean of $\bar{z}'_{ULM}$, which is 282.8. Of 53 state standards, for about 89%, $\bar{z}_{ULM}$ is lower than y_{WAM}, and for about 77%, y_{WAM} is lower than $\bar{z}'_{ULM}$. Typically, y_{WAM} lies between $\bar{z}_{ULM}$ and $\bar{z}'_{ULM}$.

Although y_{WAM} is usually larger than $\bar{z}_{ULM}$, in some cases, $\bar{z}_{ULM}$ is higher. For example, for G4 of 2000 mathematics in North Carolina (*inconsistent mastery*), $\bar{z}_{ULM}$ is about 6.6 points higher than y_{WAM}. This is the largest positive discrepancy among all jurisdictions. For G8 of 2000 mathematics, the $\bar{z}_{ULM}$ of Hawaii (*HI Stanine 5*) is 3.4 points higher than y_{WAM}, which is the largest positive discrepancy among all jurisdictions.

For G8 of 2000 mathematics, on average, y_{WAM} is 2.1 points higher than $\bar{z}_{ULM}$ but 0.6 points lower than the mean of $\bar{z}'_{ULM}$, which is 282.8. Of 53 state standards, for about 89%, $\bar{z}_{ULM}$ is lower than y_{WAM}, and for about 77%, y_{WAM} is lower than $\bar{z}'_{ULM}$. Typically, y_{WAM} lies between $\bar{z}_{ULM}$ and $\bar{z}'_{ULM}$.

Although y_{WAM} is usually larger than $\bar{z}_{ULM}$, in some cases, $\bar{z}_{ULM}$ is higher. For example, for G4 of 2000 mathematics in North Carolina (*inconsistent mastery*), $\bar{z}_{ULM}$ is about 6.6 points higher than y_{WAM}. This is the largest positive discrepancy among all jurisdictions. For G8 of 2000 mathematics, the $\bar{z}_{ULM}$ of Hawaii (*HI Stanine 5*) is 3.4 points higher than y_{WAM}, which is the largest positive discrepancy among all jurisdictions.

Table 5

G4 2000 Math: The Unweighted and Weighted NAEP Equivalents to the State Standards

State & standard	Scale scores $\bar{z}_{ULM}$ (1)	Scale scores (ULM w/o FPE) $\bar{z}'_{ULM}$ (2)	Scale scores y_{WAM} (3)	(1)-(3) (4)	SD of $\{z_l\}$ (5)	Variance of (1) by $v_{SRS}(\bar{z}_{ULM})$ (6)	Jackknifed variance of (3) $v_J(y_{WAM})$ (7)	Measurement error of (3) $(1+M^{-1})B$ (8)	Total variance of (3) $v_T(y_{WAM})$ (9)
AR benchmark	229.2	229.6	229.3	-0.1	8.5	0.6	1.8	0.2	1.9
CA PR25	194.7	198.7	192.3	2.4	13.8	2.4	2.9	0.9	3.8
CA PR50[a]	213.8	217.9	212.0	1.8	11.1	1.6	2.6	0.2	2.8
CA PR75	232.1	236.0	230.9	1.2	10.2	1.3	4.8	0.3	5.0
CT goal	224.6	227.2	225.4	-0.8	9.2	0.7	2.2	0.0	2.2
GA meets[a]	209.4	210.4	209.0	0.4	9.8	0.9	1.7	0.1	1.8
GA exceeds	254.6	254.1	257.0	-2.4	8.7	0.7	2.6	0.4	3.0
KS basic	204.7	205.7	204.6	0.1	13.3	2.1	4.1	1.1	5.2
KS satisfactory	226.9	228.5	228.5	-1.6	11.1	1.5	2.8	1.4	4.1
KS proficient[a]	244.3	245.2	244.4	-0.1	8.2	0.8	1.2	0.2	1.4
KS advanced	263.9	263.9	264.5	-0.6	9.3	1.0	1.3	0.2	1.5
LA appro. basic	199.6	202.3	200.9	-1.3	11.9	1.2	2.1	0.1	2.3
LA basic	217.5	219.0	217.9	-0.4	9.9	0.8	1.2	0.2	1.3
LA proficient[a]	249.9	250.3	250.8	-0.9	9.8	0.8	0.7	0.9	1.6
LA advanced	269.2	261.8	272.6	-3.4	7.8	0.5	3.9	1.1	4.9
MA pass	207.6	211.5	207.3	0.3	14.5	1.8	5.2	0.5	5.7
MA proficient[a]	242.2	243.1	241.9	0.3	8.3	0.6	0.8	0.0	0.8
MA advanced	265.1	264.8	265.3	-0.2	7.8	0.5	1.1	0.3	1.4
ME partially meets	212.8	215.6	215.0	-2.2	13.3	1.2	0.7	0.4	1.1
ME meets[a]	248.6	249.8	248.2	0.4	9.1	0.5	0.9	0.5	1.3

(Table continues)

Table 5 (continued)

State & standard	Scale scores $\overline{z}_{ULM}$ (1)	Scale scores (ULM w/o FPE) $\overline{z}'_{ULM}$ (2)	Scale scores y_{WAM} (3)	(1)-(3) (4)	SD of $\{z_l\}$ (5)	Variance of (1) by $v_{SRS}(\overline{z}_{ULM})$ (6)	Jackknifed variance of (3) $v_J(y_{WAM})$ (7)	Measurement error of (3) $(1+M^{-1})B$ (8)	Total variance of (3) $v_T(y_{WAM})$ (9)
ME exceeds	277.7	272.3	282.9	-5.2	8.8	0.5	2.4	0.4	2.8
MI moderate	178.6	189.8	178.7	-0.1	22.9	6.0	14.4	2.7	17.1
MI satisfactory[a]	205.4	211.9	207.4	-2.0	19.6	4.4	5.3	0.2	5.6
MO progressing	174.1	180.7	170.0	4.1	19.0	3.3	21.3	0.5	21.8
MO near proficient	205.4	207.1	207.3	-1.9	14.1	1.8	2.1	0.6	2.6
MO proficient[a]	237.0	237.9	238.5	-1.5	12.8	1.5	1.2	0.3	1.5
MO advanced	262.9	263.4	266.1	-3.2	12.3	1.4	2.9	0.5	3.4
NC inconsist mastery	173.6	178.8	167.0	6.6	24.8	5.2	33.9	1.2	35.0
NC consist mastery[a]	203.1	203.2	202.7	0.4	10.7	1.0	2.2	1.0	3.1
NC superior	238.5	238.9	237.6	0.9	7.7	0.5	1.2	0.1	1.2
NE emerging	192.1	196.3	187.8	4.3	15.6	12.8	35.1	9.3	44.3[c]
NE proficient[a]	216.7	217.6	215.8	0.9	15.4	12.5	33.4	3.6	37.0
NE advanced	236.6	239.2	237.8	-1.2	14.2	10.6	32.8	1.0	33.7
NY needs improvement	181.5	191.3	186.0	-4.5	18.0	7.9	4.7	1.5	6.1
NY meets[a]	214.8	217.4	216.3	-1.5	9.8	2.3	8.8	0.0	8.8
NY exceeds	252.8	253.5	252.4	0.4	9.4	2.1	2.1	0.4	2.5
OH pass[a]	234.6	237.1	236.3	-1.7	9.6	1.1	2.1	0.2	2.3
RI proficient[a]	250.5	250.0	250.6	-0.1	11.0	0.5	1.0	0.1	1.0

(Table continues)

Table 5 (continued)

State & standard	Scale scores $\overline{z}_{ULM}$ (1)	Scale scores (ULM w/o FPE) $\overline{z}'_{ULM}$ (2)	Scale scores y_{WAM} (3)	(1)-(3) (4)	SD of $\{z_l\}$ (5)	Variance of (1) by $v_{SRS}(\overline{z}_{ULM})$ (6)	Jackknifed variance of (3) $v_J(y_{WAM})$ (7)	Measurement error of (3) $(1+M^{-1})B$ (8)	Total variance of (3) $v_T(y_{WAM})$ (9)
SC basic	212.0	213.8	211.9	0.1	10.7	0.9	1.9	0.9	2.8
SC proficient[a]	243.4	244.6	243.9	-0.5	9.4	0.7	1.4	0.4	1.8
SC advanced	262.3	262.4	264.6	-2.3	11.3	1.0	1.0	0.4	1.4
TX pass[a]	194.8	203.0	200.6	-5.8	23.7	5.5	1.5	1.0	2.5
VT meets[a]	216.4	219.4	218.4	-2.0	15.6	2.9	4.1	0.4	4.5
WY partial proficient	221.9	222.7	221.3	0.6	12.6	0.9	1.3	1.5	2.8
WY proficient[a]	246.9	248.0	246.8	0.1	9.7	0.6	1.8	0.5	2.3
WY advanced	268.5	270.1	271.5	-3.0	9.6	0.5	7.9	1.0	8.9

Note. WAM = weighted aggregate mapping, ULM = unweighted local mapping.

[a] The state standard of proficiency. [b] The estimate for the Missouri's *MO advanced* standard is updated. The difference between ULM estimates (1) and WAM estimates (3) is also updated. But the calculation of its standard deviation was based on the old estimate of 329.8. [c] Note that the variance estimates for Nebraska are disturbingly large. A checking of the analysis revealed that the data set employed was a 20% subset of the full data set. The reduced sample size accounts for the tabled results

Table 6

G8 2000 Math: The Unweighted and Weighted NAEP Equivalents to the State Standards

State & standard	Scale scores $\bar{z}_{ULM}$ (1)	Scale scores (ULM w/o FPE) $\bar{z}'_{ULM}$ (2)	Scale scores y_{WAM} (3)	(1)-(3) (4)	SD of $\{z_i\}$ (5)	Variance of (1) by $v_{SRS}(\bar{z}_{ULM})$ (6)	Jackknifed variance of (3) $v_J(y_{WAM})$ (7)	Measurement error of (3) $(1+M^{-1})B$ (8)	Total variance of (3) $v_T(y_{WAM})$ (9)
AZ approach	264.7	270.7	267.7	-3.0	11.3	1.4	2.7	0.4	3.1
AZ meets[a]	301.4	304.6	302.7	-1.3	11.4	1.4	4.4	0.9	5.3
AZ exceeds	320.0	321.5	326.0	-6.0	14.6	2.3	0.8	2.9	3.7
CA PR25	235.9	244.4	239.3	-3.4	15.5	3.1	7.7	0.7	8.5
CA PR50[a]	264.8	268.3	263.9	0.9	10.5	1.4	5.6	0.3	5.9
CA PR75	290.3	292.3	291.2	-0.9	10.5	1.4	4.7	0.2	5.0
CT at goal	273.0	276.1	275.4	-2.4	11.4	0.7	1.5	0.1	1.5
GA meets[a]	261.4	262.9	261.9	-0.5	11.8	1.1	2.4	0.2	2.6
GA exceeds	307.7	307.6	309.4	-1.7	10.9	0.9	5.5	0.3	5.8
HI Stanine 5	257.6	258.9	254.2	3.4	10.2	0.2	1.7	0.3	2.0
IL meets[a]	277.6	282.2	279.8	-2.2	7.6	0.7	1.5	0.6	2.2
IL exceeds	314.6	313.5	316.5	-1.9	8.6	0.9	3.6	0.5	4.1
IN meets[a]	267.1	271.4	270.1	-3.0	7.4	0.6	1.4	0.8	2.3
LA approach basic	238.5	243.9	242.5	-4.0	12.5	1.2	1.4	0.2	1.6
LA basic	260.7	263.8	261.8	-1.1	11.8	1.1	3.4	0.1	3.5
LA proficient[a]	307.9	304.0	306.8	1.1	10.6	0.9	3.6	1.0	4.6
LA advanced	318.8	311.3	321.1	-2.3	7.4	0.4	8.4	0.6	8.9
MA pass	269.3	273.6	270.5	-1.2	9.9	0.8	1.6	0.3	1.9
MA proficient[a]	293.6	296.8	295.0	-1.4	9.3	0.7	1.5	0.3	1.8
MA advanced	319.9	321.0	322.2	-2.3	7.8	0.5	1.3	0.0	1.3
MD satisfactory[a]	271.1	276.6	273.1	-2.0	11.9	0.9	2.4	0.2	2.6
MD excellent	313.7	312.6	314.2	-0.5	12.7	1.0	1.8	0.1	1.9

(Table continues)

Table 6 (continued)

State & standard	Scale scores $\bar{z}_{ULM}$ (1)	Scale scores (ULM w/o FPE) $\bar{z}'_{ULM}$ (2)	Scale scores y_{WAM} (3)	(1)-(3) (4)	SD of $\{z_l\}$ (5)	Variance of (1) by $v_{SRS}(\bar{z}_{ULM})$ (6)	Jackknifed variance of (3) $v_J(y_{WAM})$ (7)	Measurement error of (3) $(1+M^{-1})B$ (8)	Total variance of (3) $v_T(y_{WAM})$ (9)
ME partial	273.7	275.4	274.7	-1.0	9.6	0.7	1.4	0.4	1.8
ME meets[a]	307.3	308.7	308.6	-1.3	10.4	0.8	0.8	1.4	2.2
ME exceeds	349.9	337.6	353.2	-3.3	15.2	1.8	4.3	2.7	7.0
MN pass[a]	267.3	272.9	269.8	-2.5	12.6	2.3	4.4	1.0	5.3
MO progressing	244.7	254.1	249.3	-4.6	16.8	2.4	3.7	0.4	4.1
MO near proficient	276.7	282.6	281.0	-4.3	11.6	1.1	1.0	0.1	1.0
MO proficient[a]	303.3	306.7	308.3	-5.0	10.2	0.9	1.1	0.2	1.3
MO advanced	334 b	325.0	341.3	-7.3	12.1	1.2	1.6	1.1	2.7
NC inconsist mastery	208.2	219.6	211.7	-3.5	22.2	4.0	10.3	2.5	12.7
NC consist mastery[a]	244.1	247.4	245.3	-1.2	13.8	1.5	1.8	0.5	2.3
NC superior	282.4	283.7	283.5	-1.1	9.3	6.5	2.1	0.2	2.3
NV PR25[a]	237.0	246.8	242.8	-5.8	16.4	1.5	0.8	1.0	1.8
NV PR75	290.5	294.5	292.6	-2.1	7.2	0.3	0.7	0.3	1.1
NY needs improve.	238.6	250.9	247.4	-8.8	15.0	4.7	15.0	2.8	17.8
NY meets[a]	282.3	286.1	283.4	-1.1	7.7	1.2	8.1	0.5	8.7
NY exceeds	325.6	323.3	326.1	-0.5	9.6	1.9	5.2	1.5	6.8
OK little knowledge	222.6	234.1	231.0	-8.4	20.0	2.9	2.6	0.2	2.8
OK satisfactory[a]	250.7	256.8	254.1	-3.4	11.2	0.9	2.7	0.5	3.2
OK advanced	303.6	306.9	306.1	-2.5	8.6	0.5	1.8	0.2	1.9
OR meets[a]	276.3	278.4	277.7	-1.4	9.7	0.9	3.4	0.5	3.9
OR exceeds	299.7	300.7	300.5	-0.8	8.3	0.7	2.3	0.3	2.6
RI meets[a]	291.3	297.0	293.0	-1.7	10.6	0.1	0.5	0.5	0.9

(Table continues)

Table 6 (continued)

State & standard	Scale scores $\bar{z}_{ULM}$ (1)	Scale scores (ULM w/o FPE) $\bar{z}'_{ULM}$ (2)	Scale scores y_{WAM} (3)	(1)-(3) (4)	SD of $\{z_i\}$ (5)	Variance of (1) by $v_{SRS}(\bar{z}_{ULM})$ (6)	Jackknifed variance of (3) $v_J(y_{WAM})$ (7)	Measurement error of (3) $(1+M^{-1})B$ (8)	Total variance of (3) $v_T(y_{WAM})$ (9)
SC basic	252.1	257.4	253.9	-1.8	9.1	0.6	1.1	0.6	1.7
SC proficient	295.1	297.2	296.4	-1.3	8.6	0.5	1.8	0.3	2.1
SC advanced[a]	318.6	317.6	319.5	-0.9	11.6	0.9	1.2	0.3	1.5
TX pass[a]	219.5	234.6	232.0	-12.5	24.5	5.4	8.4	1.0	9.4
VA pass[a]	254.9	267.0	265.4	-10.5	15.5	1.7	2.7	0.7	3.4
VT meets	272 c	270.1	268.8	3.2	8.5	0.4	3.4	0.7	4.2
WY partial	260.5	262.0	261.3	-0.8	6.7	0.2	1.1	0.5	1.6
WY proficient[a]	293.2	294.1	292.0	1.2	7.1	0.2	0.9	0.1	1.0
WY advanced	321.6	319.6	319.1	2.5	7.8	0.2	2.1	1.7	3.8

Note. WAM = weighted aggregate mapping, ULM = unweighted local mapping.

[a] State standard of proficiency. [b] The estimate for Missouri's MO advanced standard is updated. The difference between ULM estimates (1) and WAM estimates (3) is also updated. But the calculation of its standard deviation was based on the old estimate of 329.8. [c] The estimate for the Vermont Grade 8 percent meeting the standard is updated by using the most recent version of the 2000 database. The difference between ULM estimates (1) and WAM estimates (3) is also updated. But the calculation of its standard deviation was based on the old estimate of 285.7.

4. Estimation of Variances of the NAEP Equivalents to the State Standards

4.1 Variance Estimation of Simple Average of School Scores

Inasmuch as NAEP estimates are based on a sample from a population, they are subject to uncertainty due to sampling. Because of the effects of cluster selection (students within schools) and of the effects of nonresponse and poststratification adjustments, observations made on different students cannot be assumed to be independent of each other. Furthermore, to account for the differential probabilities of selection, each student has an associated sampling weight, which should be used in the computation of any statistic, and which is itself subject to sampling variability.

Ignoring the effects of a complex sample design usually results in underestimating the true sampling variability. If the statistic does not use sampling weights (e.g., the simple average $\overline{z}_{ULM}$), it implicitly treats schools as if the data were collected by simple random sampling. Following this logic, an estimate of the variance of $\overline{z}_{ULM}$,including a finite population correction, yields the following variance estimate

$$v_{SRS}\left(\overline{z}_{ULM}\right) = \frac{1-f}{n\left(n-1\right)}\sum_{k=1}^{n}\left(z_k - \overline{z}_{ULM}\right)^2 ,$$

where *n* is the number of schools in a sample, *f* is the fraction of schools selected, and z_k is the NAEP equivalent for school k. Note that McLaughlin (2000) neither employed the finite population correction nor accounted for measurement error.

4.2 The Variances of Estimated NAEP Equivalents to State Standards

To complete the presentation of the methodology proposed here, it is necessary to provide an appropriate estimate of variance. Our approach was developed based on the standard NAEP methods for the estimation of the variances of reporting statistics (Allen, Donoghue, & Schoeps, 2001). The total variance of the estimate of the NAEP equivalent of a state standard consists of two independent 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 NAEP score. The sampling error was estimated by the jackknife replicate resampling (JRR) procedure applied both to schools (for the state data) and to students (for the NAEP

data). The measurement error was estimated by utilizing the variability among the plausible values generated for each assessed student.

4.2.1 The NAEP Jackknife Replicate Resampling Approach

The JRR procedure for NAEP involves the formation of a large number of strata, typically consisting of pairs of schools. In NAEP, there are usually 62 strata. For the *j*th replicate, one school in the *j*th stratum is randomly deleted, and an appropriate set of weights is computed. The calculation of the 62 jackknife replicate weights for NAEP state samples can be found in the NAEP 1998 Technical Report (Allen et al., 2001) and in Wolter (1985).

To implement the JRR in this study, we needed not only the jackknife replicate weights for students but also the jackknife replicate weights for schools, which are formed by the same procedure described in Section 3.1. For the *j*th replicate, we applied the *j*th jackknife replicate weights for schools to estimate the corresponding proportion of students meeting the standard, $\overline{p}_{w,(j)}$. Then we mapped $\overline{p}_{w,(j)}$ to the NAEP scale and found the point $y_{WAM,(j)}$, the $(1-\overline{p}_{w,(j)})$th quantile of the distribution of NAEP scores based on that same replicate and employing the corresponding replicate weights for students. Finally, the variance of the estimate y_{WAM} that is due to sampling was estimated by:

$$v_J\left(y_{WAM}\right) = \sum_{j=1}^{62}\left(y_{WAM,(j)} - y_{WAM}\right)^2 .$$

4.2.2 Estimation of the Imputation Errors and Total Variances

The measurement error component was estimated by carrying out the estimation procedure outlined in Section 2 for each of the $M = 5$ sets of plausible values. Let the NAEP equivalent of a state standard estimated by *m*th set of plausible values be $y_{WAM,m}$, $m = 1, \ldots, M$, and denote the mean of $y_{WAM,m}$ by $\overline{y}_{WAM,\cdot}$. Finally, let

$$B = \sum_{m=1}^{M}\frac{\left(y_{WAM,m} - \overline{y}_{WAM,\cdot}\right)^2}{M-1}.$$

Then the total variance is estimated by

$$v_T\left(y_{WAM}\right) = v_J\left(y_{WAM}\right) + \left(1 + M^{-1}\right)B,$$

where $\left(1 + M^{-1}\right)$ is a finite population correction factor. The estimation process mimicked that of operational NAEP: The calculation of $v_J\left(y_{WAM}\right)$ was based on the first plausible value, and the estimation of *B* was based on all five plausible values. For details see the NAEP 1998 Technical Report (Allen, et al., 2001).

4.3 Evaluation of the Variance Estimates

In Tables 5 and 6, for G4 and G8 of 2000 mathematics respectively, column 6 displays the variance of $\overline{z}_{ULM}$, obtained by application of the formula in Section 4.1, while columns 7, 8, and 9 display the error variance due to sampling, the error variance due to measurement uncertainty, and the total error variance of y_{WAM}, respectively. For 2002 state reading tests, we only computed the error variance due to sampling, the variance due to measurement error, and the total variance of y_{WAM}. Tables 7 and 8 contain these results for G4 and G8 respectively.

Returning to Tables 5 and 6, we can first compare the jackknifed variances, $v_J\left(y_{WAM}\right)$, of column 7 with the variances in column 6, obtained by use of the formula $v_{SRS}\left(\overline{z}_{ULM}\right)$. On average, for G4 and G8 of 2000 mathematics, the jackknifed variances are 5.9 and 3.2, while the corresponding averages of $v_{SRS}\left(\overline{z}_{ULM}\right)$, are 2.4 and 1.4. Clearly the effect of complex sampling on the variance of estimates is substantial, and $v_{SRS}\left(\overline{z}_{ULM}\right)$ underestimates the true sampling variability. For G4 and G8 of 2002 reading, the average jackknifed variances are 3.7 and 1.6 respectively. The average measurement errors are 0.82 and 0.68 for G4 and G8 of 2000 mathematics (representing 12% and 18% of the total variance, respectively), and 0.75 and 0.58 for G4 and G8 of 2002 reading (representing 17% and 26% of the total variance, respectively). The last figure is rather higher than the others, even though the measurement error for G8 of 2002 reading is smallest. The explanation is that the sampling error for G8 of 2002 reading is relatively small. Although the measurement errors are only a small portion of the total variance, ignoring measurement error would further underestimate the true variance of the estimators.

Table 7

G4 2002 Reading: The Weighted NAEP Equivalents to the State Standards

State & standard	Scale scores y_{WAM} (1)	Jackknifed variance of (1) $v_J(y_{WAM})$ (2)	Measurement error of (1) $(1+M^{-1})B$ (3)	Total variance of (1) $v_T(y_{WAM})$ (4)
AR proficient[a]	210.4	2.9	0.2	3.0
AR advanced	268.8	2.7	1.4	4.0
CA PR25	183.0	9.7	0.5	10.2
CA PR50[a]	209.5	8.3	0.1	8.4
CA PR75	232.3	5.9	0.9	6.8
CT Level 1	197.3	4.0	0.2	4.2
CT Level 2[a]	209.9	2.3	0.4	2.7
CT Level 3	224.0	1.3	0.5	1.9
FL Level 1	199.5	3.5	0.2	3.6
FL Level 2[a]	214.7	1.5	0.1	1.6
FL Level 3	239.1	0.9	0.1	1.0
FL Level 4	265.7	1.1	1.1	2.2
GA meets standard[a]	183.1	2.1	0.7	2.8
GA exceeds standard	224.3	1.6	0.2	1.8
ME partially meets standard	179.1	1.3	4.0	5.3
ME meets standard	226.4	1.2	0.5	1.6
ME exceeds standard	294.3	28.8[b]	4.9	33.7
MA pct passing	188.2	6.6	1.6	8.2
MA pct proficient[a]	231.5	1.1	0.2	1.2
MA pct advanced	276.5	2.0	0.5	2.5
MI pct moderate	189.3	2.4	0.5	2.9
MI pct satisfactory[a]	214.9	2.7	0.1	2.7
MS pct basic	154.3	3.7	0.9	4.5
MS pct proficient[a]	167.3	4.0	0.8	4.9
MT pct at + near proficient	180.0	4.1	1.4	5.5
MT pct at + proficient[a]	200.0	3.2	0.7	3.9
MT pct advanced	253.7	2.0	1.0	3.1
NY Level 1	170.5	3.8	0.4	4.3
NY Level 2[a]	213.6	2.6	0.2	2.9
NY Level 3	252.7	1.5	0.2	1.7

(Table continues)

Table 7 (continued)

State & standard	Scale scores y_{WAM} (1)	Jackknifed variance of (1) $v_J(y_{WAM})$ (2)	Measurement error of (1) $(1+M^{-1})B$ (3)	Total variance of (1) $v_T(y_{WAM})$ (4)
NC Level 1	167.3	2.9	1.4	4.3
NC Level 2[a]	198.0	1.6	0.4	2.0
NC Level 3	238.1	1.4	0.1	1.5
OH pct passing[a]	208.6	2.9	0.8	3.7
RI pct prof. (analysis)[a]	211.7	1.3	0.4	1.7
RI pct prof. (basic)[a]	198.5	3.2	0.2	3.3
SC pct passing	185.5	2.9	0.6	3.5
SC pct proficient[a]	231.3	0.7	0.6	1.2
SC pct advanced	280.6	1.1	1.6	2.7
TX pct passing[a]	167.9	2.7	0.1	2.8
TX pct mastering	218.5	2.8	0.2	3.0
VT pct meet basic[a]	200.1	0.8	1.6	2.4
WA Level 1	162.6	8.0	0.5	8.5
WA Level 2[a]	209.5	2.0	0.1	2.0
WA Level 3	243.5	2.8	0.1	2.9
WI pct basic	164.4	17.0	1.1	18.0
WI pct proficient[a]	192.8	9.3	1.7	11.0
WI pct advanced	250.9	1.6	1.3	2.9
WY pct partial proficient	195.1	1.4	0.7	2.2
WY pct above proficient[a]	228.1	2.2	0.1	2.3
WY pct advanced	255.4	1.1	0.1	1.2

[a] State standard of proficiency.[b] The jackknifed variance estimates for *ME exceeds standard*, 28.8, is relatively large compared with the other two standards. Since only 0.58% of Maine students meet the *exceeds standard*, the number of students at each school meeting the standard is small, and this results in large variation in jackknifed estimates. Such a number could be suppressed in reporting. To obtain better estimates, we can apply robust statistical techniques such as the Winsorized variance estimate. For its application, see the note in Table 9.

Table 8

G8 2002 Reading: The Weighted NAEP Equivalents to the State Standards

State & standard	Scale scores y_{WAM} (1)	Jackknifed variance of (1) $v_J(y_{WAM})$ (2)	Measurement error of (1) $(1+M^{-1})B$ (3)	Total variance of (1) $v_T(y_{WAM})$ (4)
AR basic	226.2	2.4	0.4	2.8
AR proficient[a]	277.5	1.1	0.7	1.8
AR advanced	315.9	2.6	1.5	4.1
CA PR25	229.0	3.9	0.7	4.6
CA PR50[a]	253.4	9.2	0.1	9.3
CA PR75	280.7	2.0	0.4	2.4
CT Level 1	231.6	4.0	0.4	4.3
CT Level 2[a]	242.8	2.5	0.6	3.1
CT Level 3	254.8	1.0	0.5	1.5
DE below standard	230.8	2.1	0.3	2.4
DE meets standard[a]	251.9	0.4	0.4	0.8
DE exceeds standard	302.8	0.3	0.2	0.5
DE distinguished	317.3	1.0	0.3	1.4
FL Level 1	240.6	3.0	1.0	3.9
FL Level 2[a]	265.3	2.6	0.3	2.9
FL Level 3	291.9	6.0	0.7	6.7
FL Level 4	317.1	1.3	5.2	6.4
GA meets standard[a]	231.3	1.7	0.4	2.2
GA exceeds standard	266.1	1.9	0.3	2.2
HI Stanine 4+	231.0	1.0	0.3	1.3
HI Stanine 5+[a]	251.3	0.5	0.4	1.0
Hi Stanine 7+	278.8	0.4	0.7	1.1
IL meets standard[a]	253.6	6.8	0.4	7.2
IL exceeds standard	302.7	7.6	2.1	9.7
IN pct at or above[a]	252.7	1.3	0.8	2.1
KS pct basic	228.8	5.1	0.9	6.0
KS pct satisfactory	259.2	2.0	0.5	2.6
KS pct proficient[a]	282.1	1.1	0.5	1.7
KS pct advanced	311.5	3.2	0.1	3.4

(Table continues)

Table 8 (continued)

State & standard	Scale scores y_{WAM} (1)	Jackknifed variance of (1) $v_J(y_{WAM})$ (2)	Measurement error of (1) $(1+M^{-1})B$ (3)	Total variance of (1) $v_T(y_{WAM})$ (4)
ME partially meets standard	234.0	1.7	1.1	2.8
ME meets standard[a]	276.7	0.9	0.2	1.1
ME exceeds standard	330.5	3.8	5.1	8.9
MD pct satisfactory[a]	281.9	4.4	0.7	5.1
MD pct excellent	317.2	13.4[b]	3.8	17.2
MS pct basic	233.3	1.1	0.2	1.3
MS pct proficient[a]	256.7	1.4	0.5	1.9
MT pct at + near proficient	239.0	4.6	2.5	7.1
MT pct at + proficient[a]	255.5	1.6	0.3	1.8
MT pct advanced	296.6	0.4	2.1	2.5
NY Level 1	215.4	3.7	2.5	6.2
NY Level 2[a]	273.3	1.0	0.8	1.8
NY Level 3	304.7	1.6	1.0	2.6
NC Level 1	194.3	8.1	3.2	11.3
NC Level 2[a]	231.3	1.9	0.6	2.6
NC Level 3	273.7	1.4	1.0	2.4
OR pct meet or exceed[a]	256.9	3.0	0.4	3.4
OR pct exceed	282.0	1.3	0.2	1.5
PA pct basic	239.1	2.1	0.2	2.4
PA pct proficient[a]	261.8	1.5	0.5	2.0
PA pct advanced	293.5	0.6	0.5	1.1
RI pct prof (analysis)[a]	283.4	0.4	0.3	0.7
SC pct passing	242.6	0.9	1.0	1.8
SC pct proficient[a]	278.9	1.0	0.4	1.4
SC pct advanced	309.9	1.2	1.1	2.3
TX pct passing	202.9	5.5	1.2	6.7
TX pct mastering[a]	258.9	2.7	0.3	3.0
VT pct meet basic	261.5	0.5	0.3	0.8

(Table continues)

Table 8 (continued)

State & standard	Scale scores y_{WAM} (1)	Jackknifed variance of (1) $v_J(y_{WAM})$ (2)	Measurement error of (1) $(1+M^{-1})B$ (3)	Total variance of (1) $v_T(y_{WAM})$ (4)
VA pct passing[a]	253.1	1.6	0.1	1.7
WY pct partial proficient	242.4	0.6	1.6	2.2
WY pct above proficient[a]	275.4	0.8	0.1	0.9
WY pct advanced	306.7	0.8	0.6	1.5

[a]State standard of proficiency. [b] Similar to the ME exceeds standard of G4 2002 reading, the jackknifed variance estimates for MD pct excellent, 13.4, is also relative large compared with other two standards. Since only 2.6% Maryland students meet the pct excellent, the number of students at each school meeting the standard is small and results in large variation in jackknifed estimates. Such number could be suppressed in reporting. To obtain better estimates, we can apply robust statistical techniques such as the Winsorized variance estimate. For its application, see the note in Table 9.

Kish (1965) defined the design effect (DEFF) as the ratio of the variance of a statistic from a complex sample to the variance of the statistic from a simple random sample of the same size. If $v_{SRS}(\bar{z}_{ULM})$ is treated as a variance estimate based on simple random sampling, the design effect for the NAEP equivalent of the 2000 state mathematics standard ranges from 2.0 to 2.5. This is consistent with the design effects for reported NAEP statistics. It shows that the complex sampling effects cannot be ignored in the calculation of variances.

The differences in the estimated variances for $\bar{z}_{ULM}$ and y_{WAM} are illustrated in Figure 2, which contains two sets of plots for G4 2000 math: (a) A plot of y_{WAM} against an estimate of its total variance, and (b) A plot of $\bar{z}_{ULM}$ against its estimated variance, using the formula $v_{SRS}(\bar{z}_{ULM})$. Nearly all the rhombus icons, representing the total variances of y_{WAM}, are located to the right of the triangle icons representing the variances of $\bar{z}_{ULM}$. Figure 3 shows the same pattern for G8 2000 mathematics. Even with the larger estimated variances, for most states, the magnitudes of the estimated standard deviations of the mapped equivalents are modest in comparison to the differences among the equivalents. Note that these variances address only one

aspect of the stability of the estimated equivalents. Other relevant evidence would be obtained by carrying out the linkage separately for different subgroups of the student population. Unfortunately, the requisite data are generally not available.

5. Findings

5.1 The State Standards for the 2000 State Mathematics Tests

Based on the analysis of the 2000 mathematics data, the results obtained through WAM show the same patterns as the results obtained through ULM. Both approaches support the credibility of the estimated NAEP equivalents to the state standards. The main finding is that the mapped NAEP scale scores, either $\overline{z}_{ULM}$ or y_{WAM}, are very strongly inversely related to the percentages of students at or above a state standard: If a state has lower percentage of students above its standard, then that standard typically maps into a higher NAEP scale score. For WAM, these findings are illustrated in Figures 4-7 for the 2000 state mathematics tests.

For example, in Figure 5,[7] based on G8 of 2000 mathematics, Maine has the highest mapped NAEP scale score. Its *exceeds the standard* category has 1.1% at or above this standard and its NAEP equivalent is 353. The second most stringent standard is Montana's *advanced* category, with 1.2% at or above this standard and a NAEP equivalent of 341. The next most stringent one is Louisiana's *advanced* category, with 2.7% at or above this standard and a mapped standard of 321. The least stringent standard is North Carolina's *inconsistent mastery* category, with 95.6% of the students meeting the standard and a mapped standard of 212. These results are consistent with those obtained by ULM.

An ordinary least squares regression line has been superimposed on Figure 5. There is relatively little scatter about the line, even at extreme values of percent above the standard. The pattern is clear: States with higher percentages above their standard tend to have a lower NAEP equivalent to that standard. The correlation in Figure 5 is -.96. It is important to recognize that the observed pattern is not a logical consequence of the methodology. Now, if one were to construct a comparable figure based on the quantiles of a single, approximately normal distribution (e.g., the national NAEP distribution for G4 2000 math), then one would obtain a straight line relationship, particularly for percents between 20 and 80. However, the data points in Figure 5 were drawn from many different states, each with its own test and distribution of test scores.

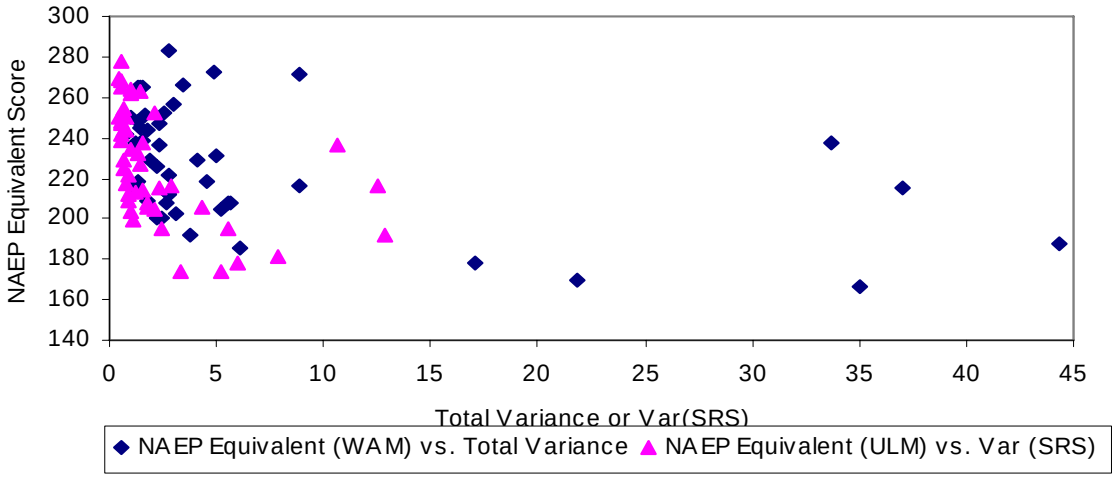

***Figure 2*. G4 2000 math (proficient): NAEP equivalent (WAM or ULM) versus variance [total variance or Var(SRS)].**

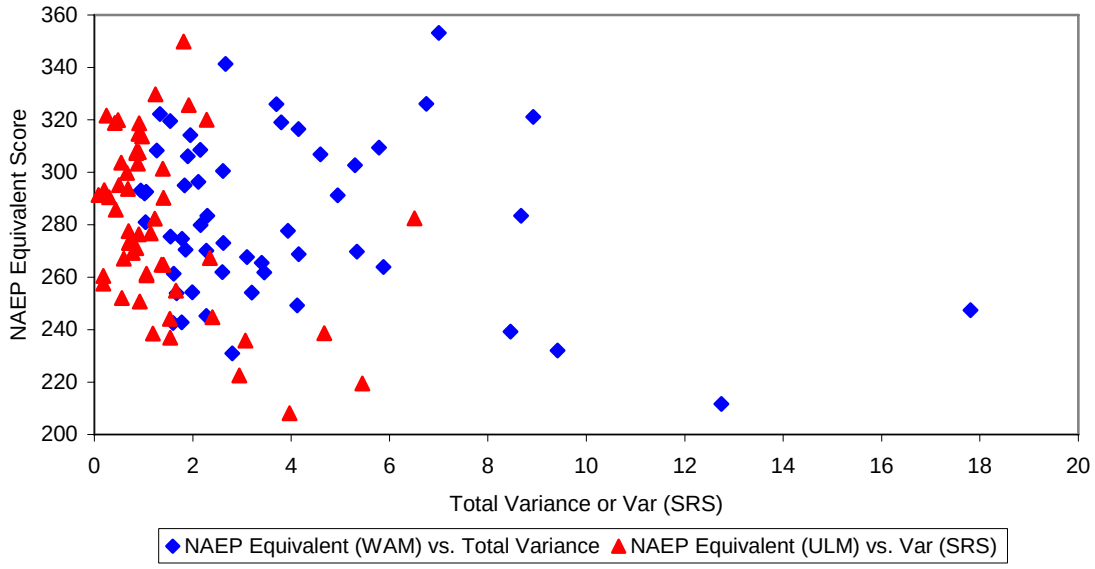

***Figure 3*. G8 2000 math (proficient): NAEP equivalent (WAM or ULM) versus variance [total variance or Var(SRS)].**

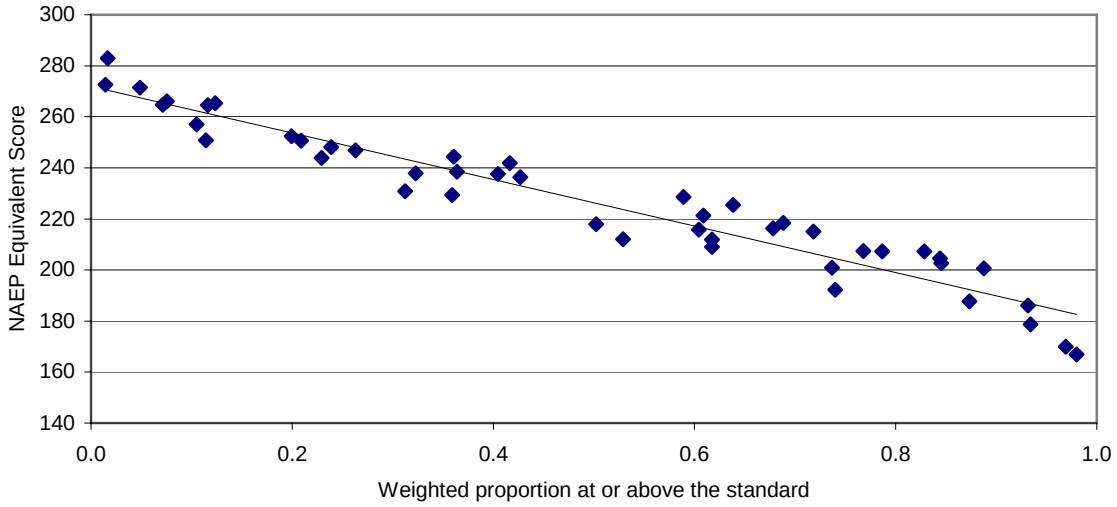

Figure 4. **G4 2000 math: NAEP equivalents to the state standards vs. proportions at or above state standards.**

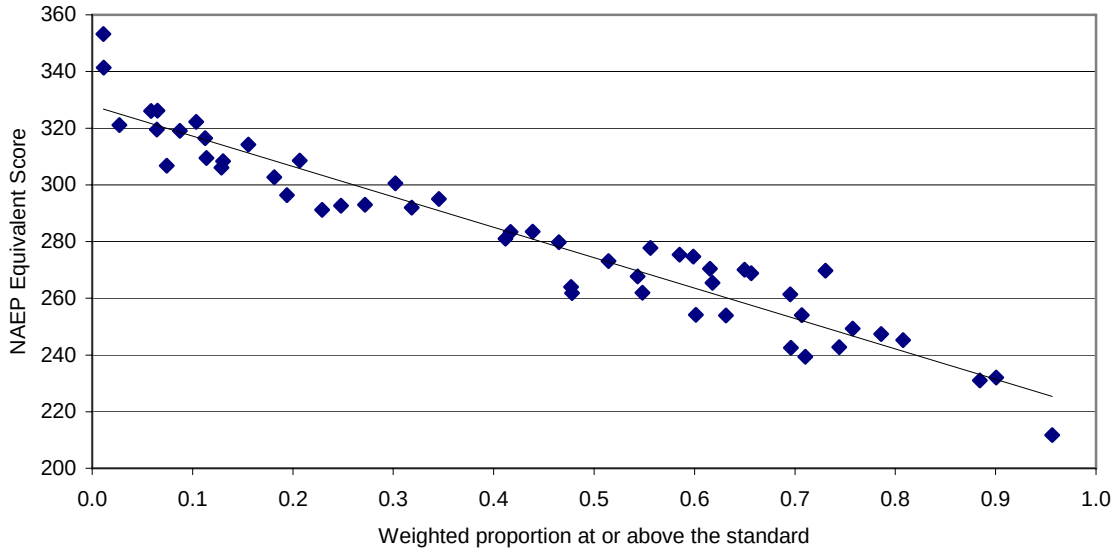

Figure 5. **G8 2000 math: NAEP equivalents to the state standards vs. proportions at or above state standards.**

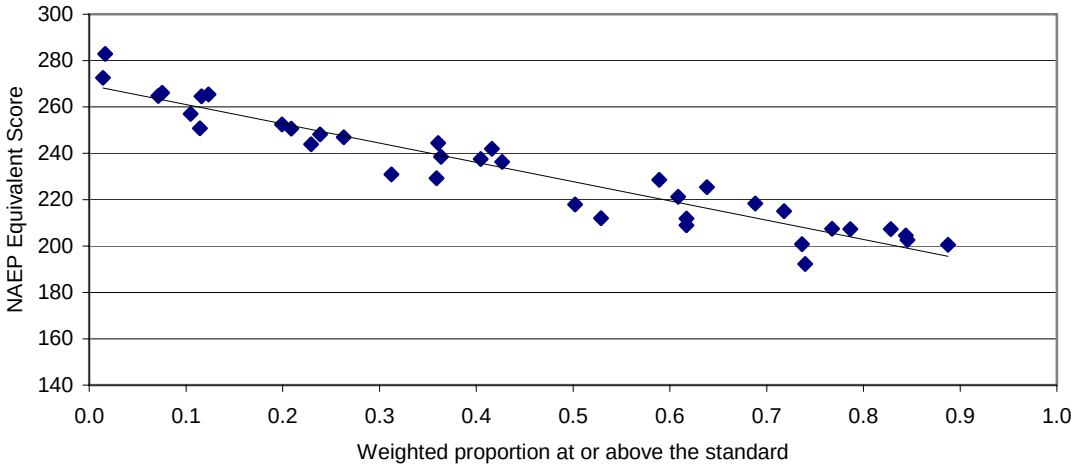

Figure 6. **G4 2000 math: NAEP equivalents to the state standards vs. proportions at or above state standards (standards with large SEs removed).**

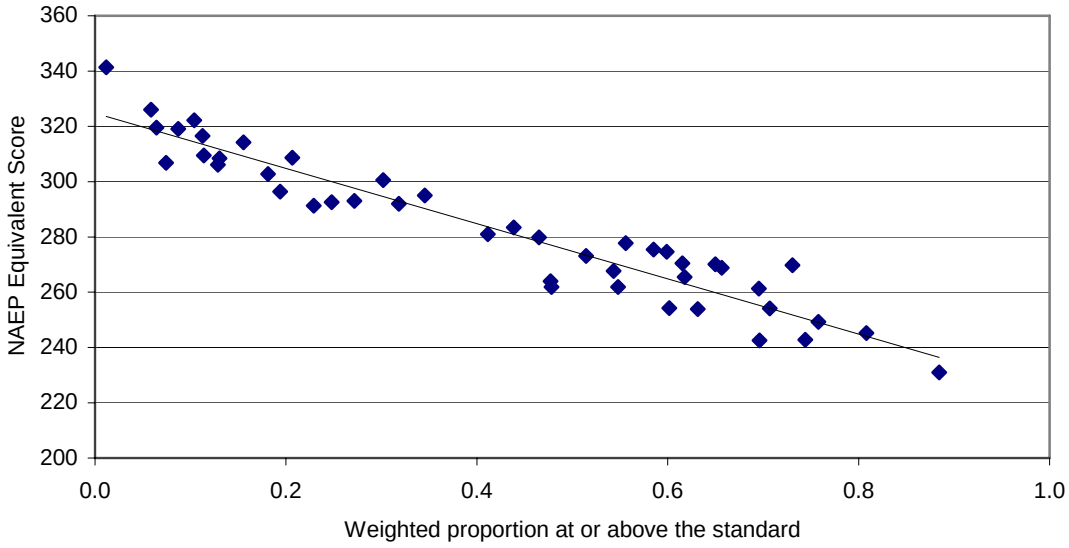

Figure 7. **G8 2000 math: NAEP equivalents to the state standards vs. proportions at or above state standards (standards with large SEs removed).**

The availability of estimated variances for mapped standards makes possible the construction of confidence intervals for the mapped standards. If the confidence bands overlap the fitted regression line, then the mapped standards can be considered credible. The confidence intervals are relatively wide because, on average, the total variances are 6.75 and 3.83 for G4 and G8 of 2000 mathematics. Typically confidence bands cover the regression line, confirming the inverse relationship between percentages meeting the standard and mapped standards. While there are reversals, they are usually within the margin of error indicated by estimated variances. For example, in Figure 5 for G8 of 2000 math, [8] Arizona has 5.9% of students above its standard of *exceeds*, which is higher than the 2.7% of Louisiana's *advanced* category. But the mapped standard for Arizona is 326, which is higher than the 321 for Louisiana. However, the standard errors of the mapped NAEP scores are 3.0 and 1.9 for Arizona and Louisiana. Therefore, the difference between 321 and 326 is not significant. At the same time, we should recognize that such reversals may be due, at least in part, to real differences in the distributions of achievement between the states.

To see if the strength of the inverse relationship is greater when points corresponding to mapped standards with large estimated variances are removed, we deleted those points with estimated variances greater than 6. Figures 6 and 7 display the resulting trimmed samples for G4 and G8 of 2000 mathematics, respectively. In comparison to Figures 4 and 5, these data show somewhat less variation about the fitted line.

Figures 8 and 9 display the plots of the mapped standards y_{WAM} against their estimated variances, for G4 and G8 of 2000 mathematics, respectively. The patterns for $\overline{z}_{ULM}$ are similar to those for y_{WAM}, although the magnitudes of the variances are different. Figures 10 and 11 are the same as Figures 8 and 9, but with the points corresponding to mapped standards with large estimated variances removed. The remaining points are labeled with the corresponding state/standard. It is evident that there are still substantial differences in how states set their achievement standards.

More surprisingly, perhaps, there appears to be a wide range of expectations for student achievement, even when only state standards for proficiency are considered. Of course, such comparisons can only be made when the standards are placed on a common scale. Figures 12 and 13 display the relationship between the NAEP equivalents and the percentages of students at or above the standard, employing only the state standards for proficiency[9] for G4 and G8 2000 mathematics, respectively. We note that for both G4 and G8, most of the NAEP equivalents are

lower than the NAEP standards for proficiency, which are 249 and 299, respectively (Braswell et al., 2001). For Grade 4, we also note that the range of NAEP equivalents is about 50 points, while for Grade 8 it is about 77 points. Such differences are certainly very large in the context of NAEP scores and indeed suggest some degree of overlap between the sets of standards for the two grades. [10] Of course, such an inference requires that the pattern of differences among the mapped equivalents on the common scale (here, the NAEP scale) can be reasonably interpreted as reflecting real differences in stringency.

To the extent that interpretation is correct, one can draw useful conclusions from Figures 12 and 13. Consider data points lying on a vertical line. These correspond to states with the same value of $\overline{p}_w$; that is, they each have the same proportion of students above their respective standard. The higher a state's point, the higher its corresponding NAEP equivalent, and we infer that it has set a more stringent standard and, therefore, that its students have demonstrated superior achievement. Now consider data points lying on a horizontal line. These correspond to states with the same NAEP equivalent. The further to the right a state's point falls, the greater its value of $\overline{p}_w$, and we infer that its students have demonstrated superior achievement. Note that in Figures 12 and 13 there is minimal vertical scatter but somewhat greater horizontal scatter (taking into account the different scales on the two axes). (To some degree this is expected, since the least squares line minimizes a function of the vertical scatter.) That there is a modest amount of horizontal scatter suggests that the observed high negative correlation is not simply an artifact of the methodology.

5.2 The State Standards for 2002 State Reading Tests

Similar to the findings for the 2000 mathematics data, the results for the 2002 reading data in Figures 14 and 15 show that the mapped NAEP scores have a strong inverse relationship with the percentages of students at or above a state standard. For G4 and G8 of 2002 reading, the average total variances are 4.44 and 3.17. Again, there are a number of mapped standards with comparatively large variances. Figures 16 and 17 correspond to Figures 14 and 15, but with the points with large variances removed. The pattern of relationships is analogous to those for mathematics.

Figures 18 and 19 display plots of the mapped standards y_{WAM} against their estimated total variances for G4 and G8 of 2002 reading. Figures 20 and 21 parallel Figures 18 and 19, but with the points corresponding to mapped standards with large estimated variances removed. The points are labeled by the corresponding state/standard.

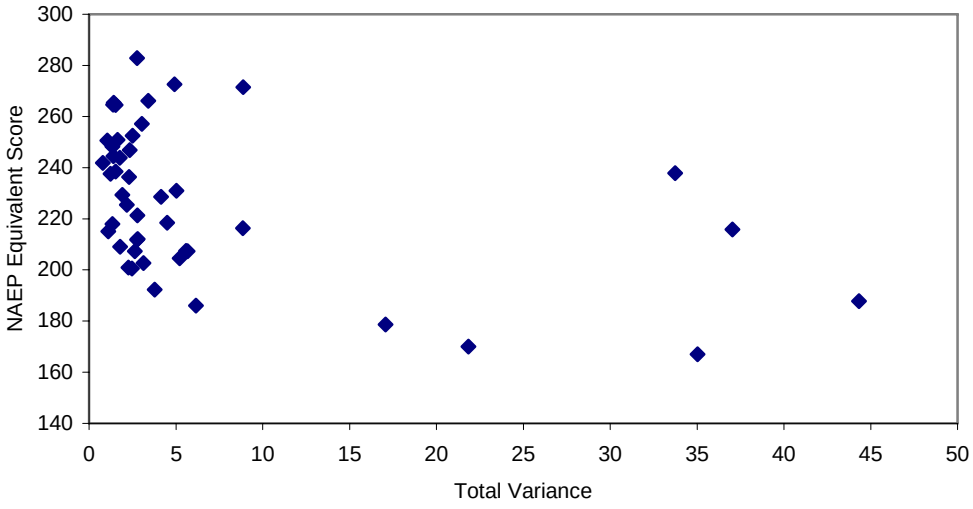

Figure 8. **G4 2000 math: NAEP equivalents to the state standards (weighted) vs. total variance.**

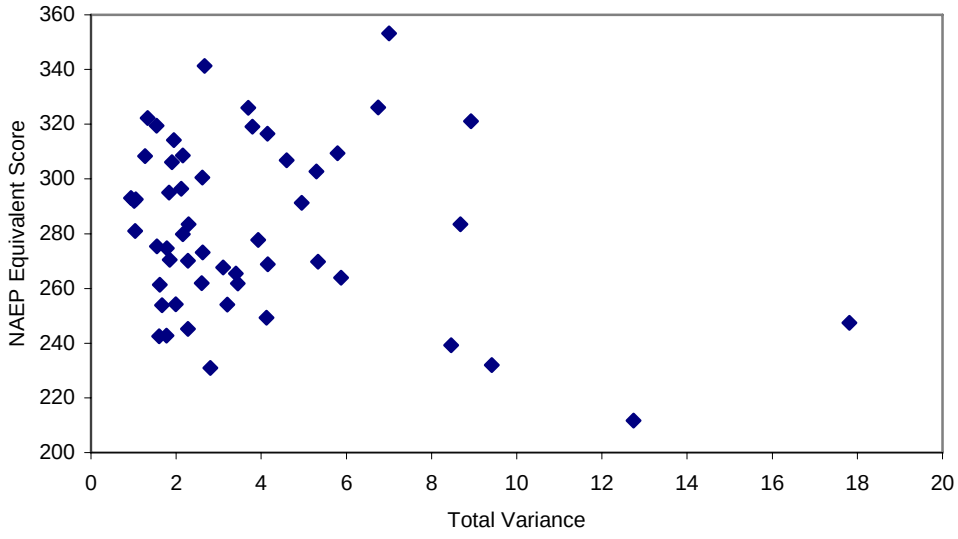

Figure 9. **G8 2000 math: NAEP equivalents to the state standards (weighted) vs. total variance.**

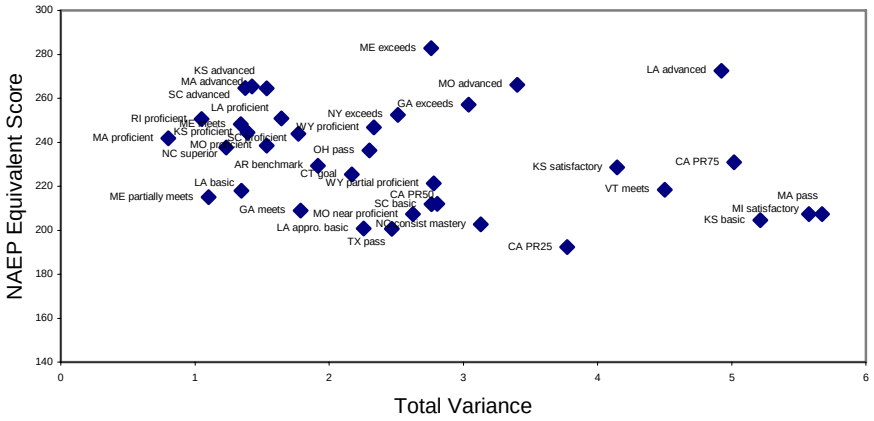

Figure 10. **G4 2000 math: NAEP equivalents to the state standards (weighted) vs. total variance (NAEP equivalents with large SEs removed).**

Note: To display the state names associated with the NAEP equivalents in Figure 10, the range of the x-axis is set differently from that in Figure 8.

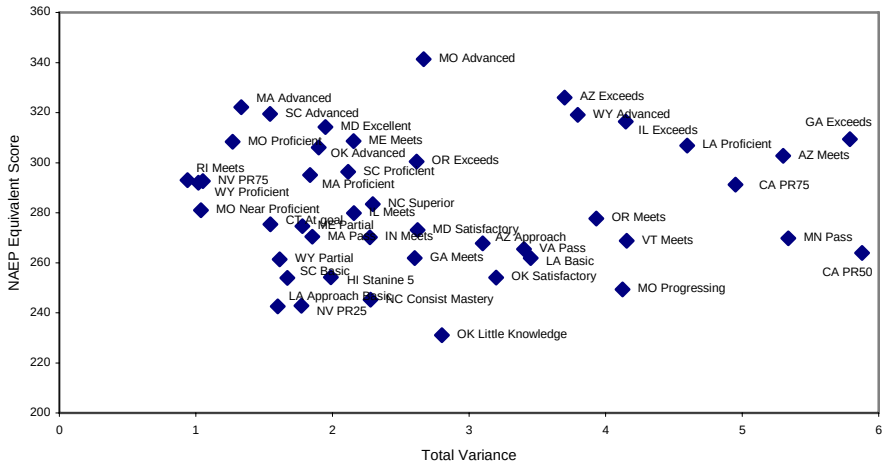

Figure 11. **G8 2000 math: NAEP equivalents to the state standards (weighted) vs. total variance (NAEP equivalents with large SEs removed).**

Note: To display the state names associated with the NAEP equivalents in Figure 11, the range of the x-axis is set differently from that in Figure 9.

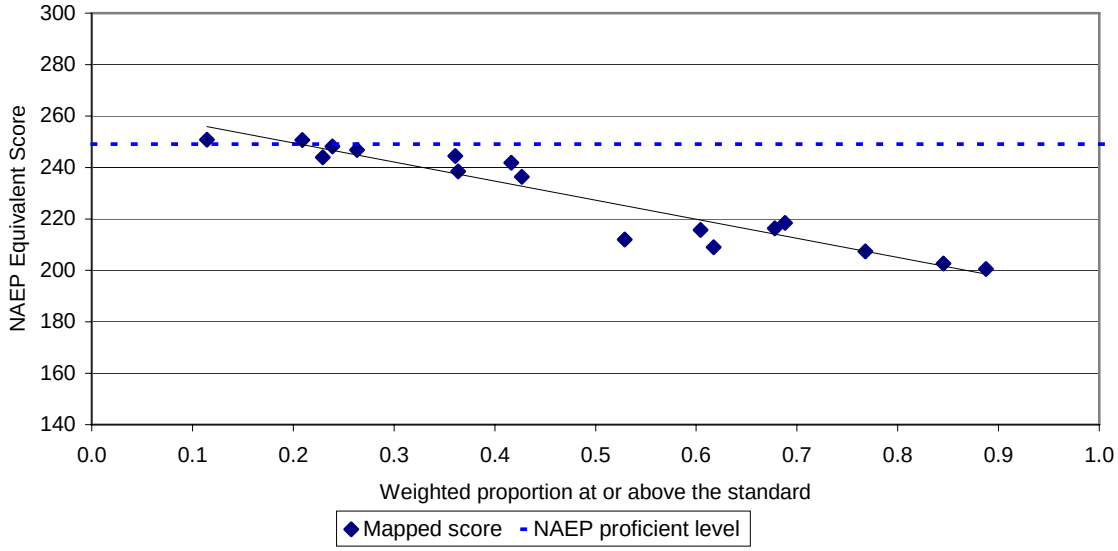

Figure 12. **G4 2000 math: NAEP equivalents to the state standards of proficient vs. proportions at or above state standards of proficient.**

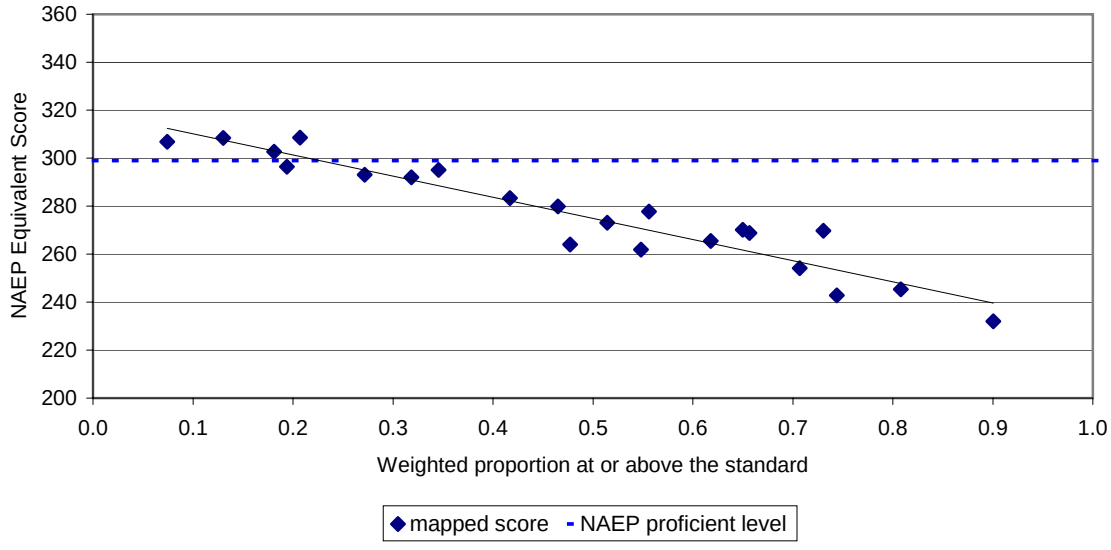

Figure 13. **G8 2000 math: NAEP equivalents to the state standards of proficient vs. proportions at or above state standards of proficient.**

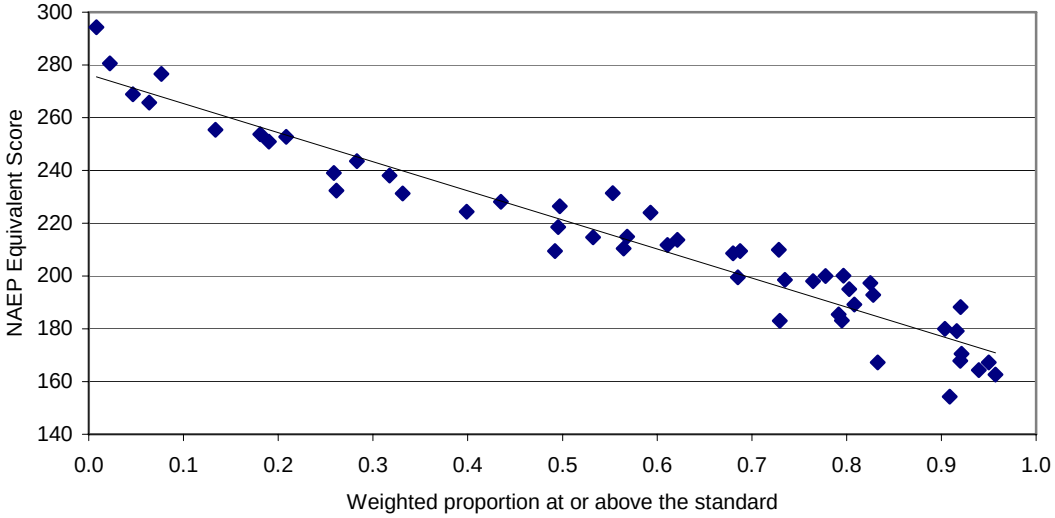

Figure 14. **G4 2002 reading: NAEP equivalents to the state standards vs. proportions at or above state standards.**

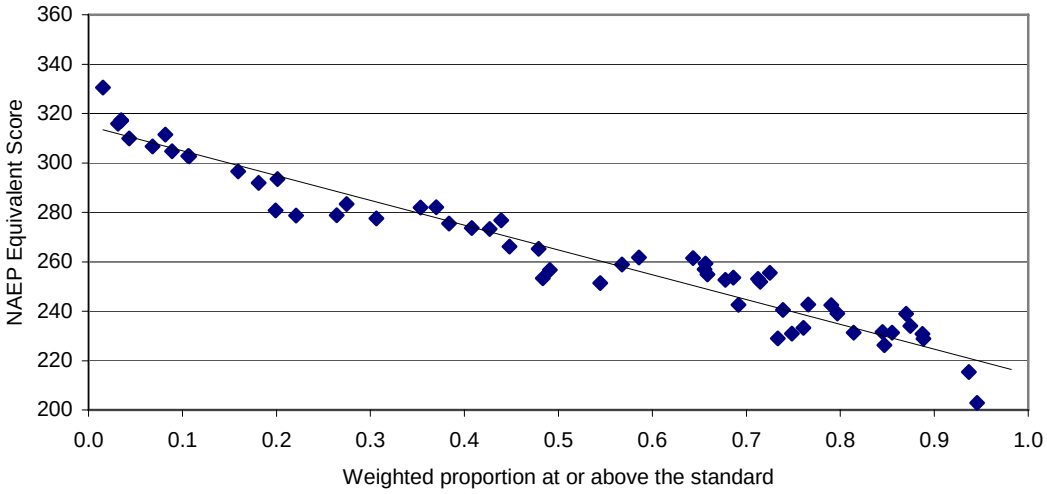

Figure 15. **G8 2002 reading: NAEP equivalents to the state standards vs. proportions at or above state standards.**

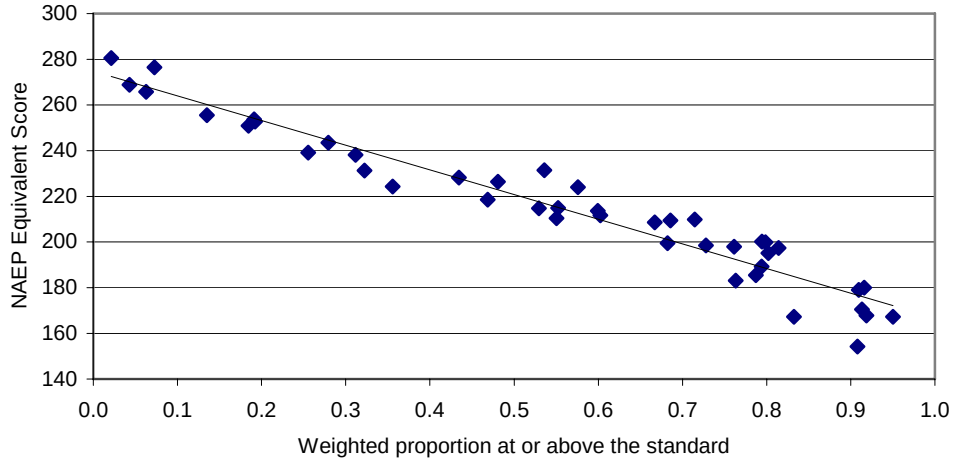

Figure 16. **G4 2002 reading: NAEP equivalents to the state standards vs. proportions at or above state standards (NAEP equivalents with large SEs removed).**

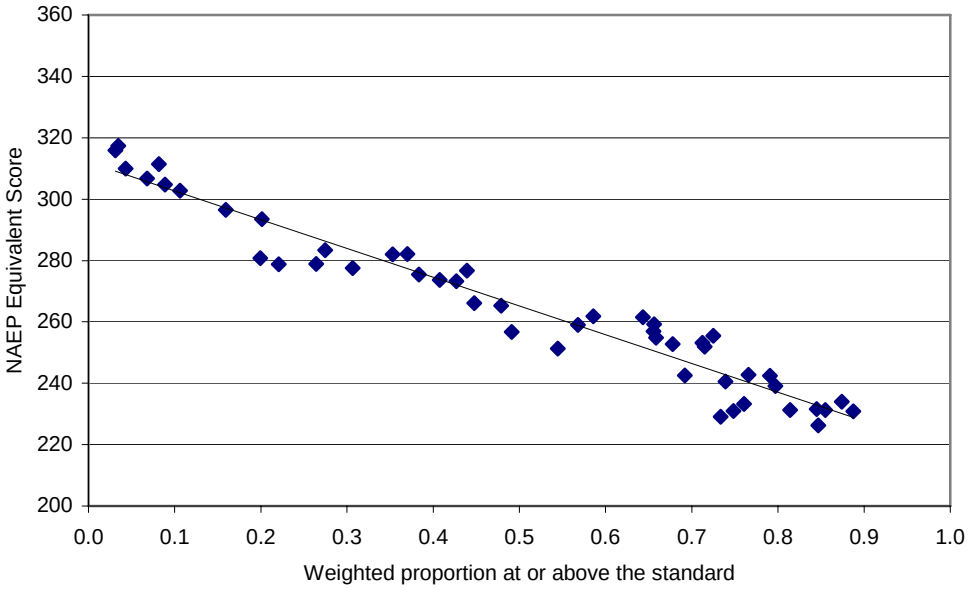

Figure 17. **G8 2002 reading: NAEP equivalents to the state standards vs. proportions at or above state standards (NAEP equivalents with large SEs removed).**

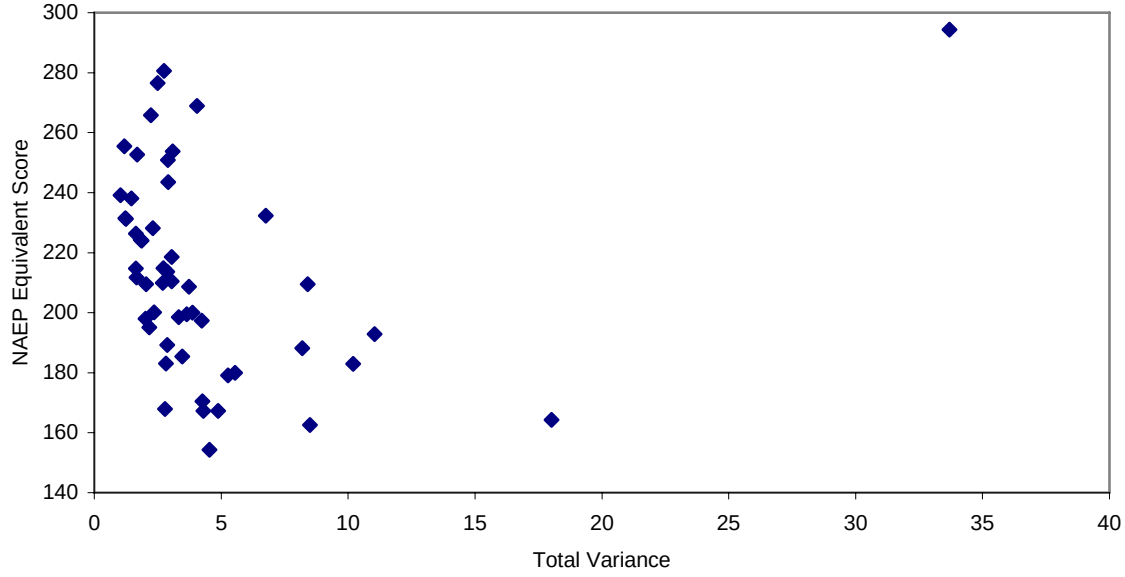

Figure 18. **G4 2002 reading: NAEP equivalents to the state standards (weighted) vs. total variance.**

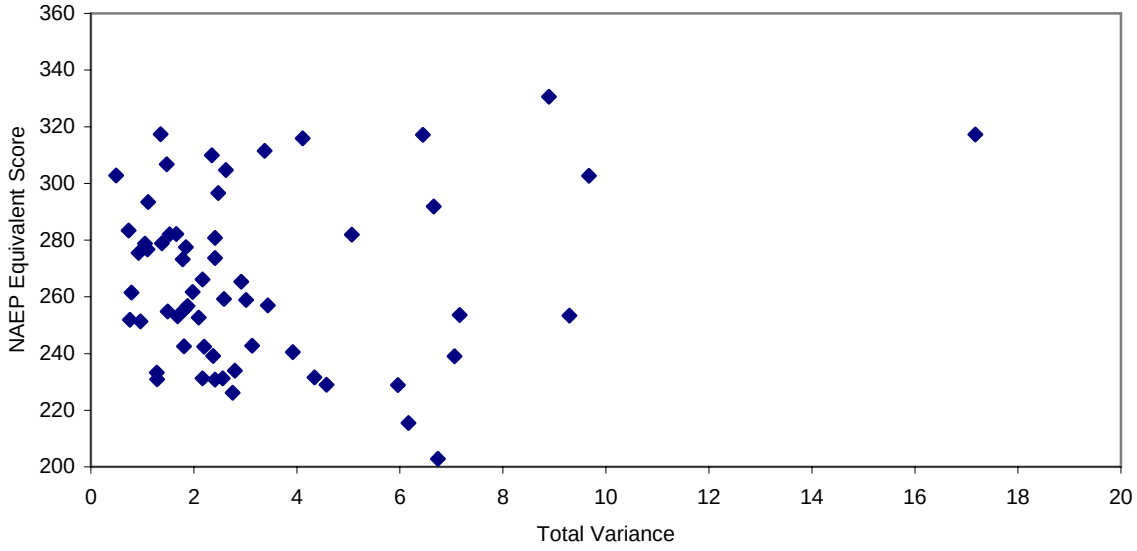

Figure 19. **G8 2002 reading: NAEP equivalents to the state standards (weighted) vs. total variance.**

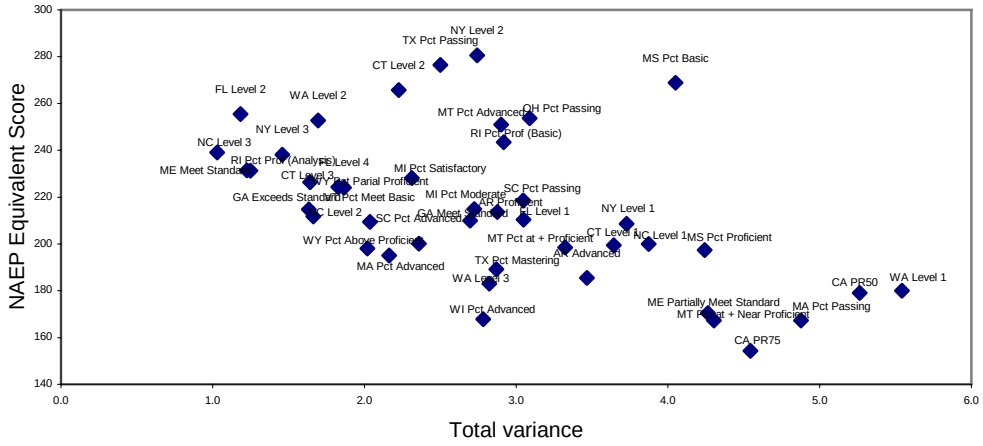

Figure 20. **G4 2002 reading: NAEP equivalents to the state standards (weighted) vs. total variance (NAEP equivalents with large SEs removed).**

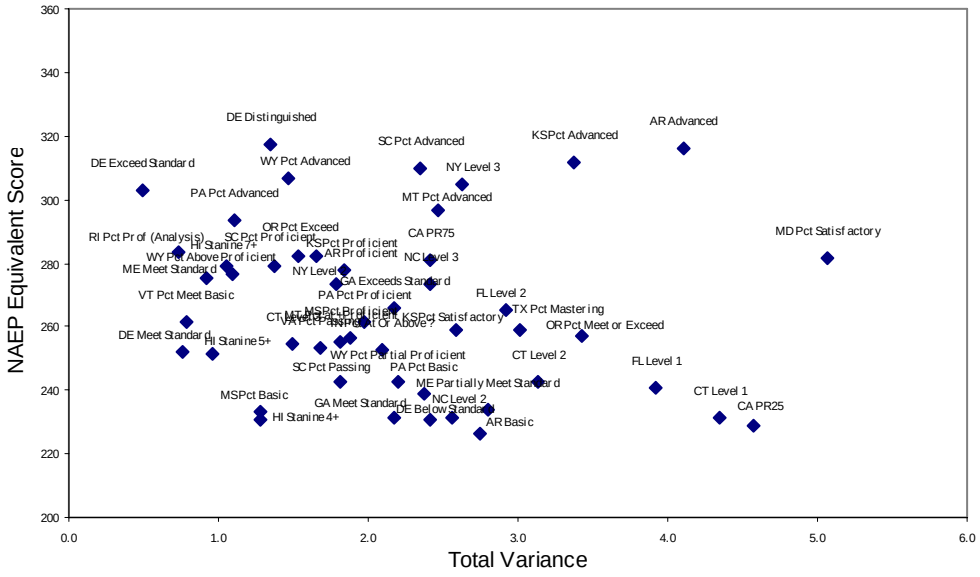

Figure 21. **G8 2002 reading: NAEP equivalents to the state standards (weighted) vs. total variance (NAEP equivalents with large SEs removed).**

Analogous to Figures 12 and 13, Figures 22 and 23 plot only those points corresponding to state standards at the proficient level, for G4 and G8 2002 reading respectively. The ranges of NAEP equivalents are about 64 and 52 points for G4 and G8 respectively. Clearly, there is a very substantial range of state standards at both Grades 4 and 8. It is striking that all of the G4 mapped proficient standards for reading are lower than the NAEP *proficient* standard at 238, and most of the mapped proficient standards for G8 are lower than the NAEP *proficient* standard at 281 (Grigg, Daane, Jin, & Campbell, 2003).

5.3 Further Considerations

Our preferred interpretation, that the variation in NAEP equivalents largely reflects differences in the stringency of states' proficiency standards, is certainly consistent with Figures 12 and 13. It is also supported by the fact that there is no, or at best, a very weak relationship between states' percent proficient and states' performance on NAEP. Figures 24 and 25 display the relevant scatter plots for Grade 4 mathematics and reading. Moreover, the heterogeneity among the NAEP equivalents is much greater than among NAEP means. [11]

To put the above results in a broader context, we carried out the mapping procedure for four different percentiles of the state test score distributions: 10^{th}, 25^{th}, 50^{th}, and 75^{th}. Figures 26 and 27 present the results for G4 2000 math and G4 2002 reading, respectively. In this setting, the dispersion among NAEP equivalents is now comparable to the dispersion among NAEP means and, moreover, the points fall very neatly along a diagonal. The correlations between the NAEP means and the NAEP equivalents to the state medians are .98 and .99 for G4 2000 math and G4 2002 reading, respectively. (Recall that the correlations between the NAEP means and the NAEP equivalents to the state standards are just .24 and .27, respectively.) For example, California had 53% of the students meeting its standard *CA PR50* in the 2000 state math test, and its NAEP mean is 213. If California were to set its proficiency standard at the 25^{th} percentile, holding all else constant, the proportion of students meeting the standard would be greater. The point corresponding to California in Figure 26 will move horizontally and settle near the diagonal line marking the NAEP equivalents for the 25^{th} percentile. Thus, Figures 26 and 27 can serve as baselines against which to judge the observed results for the state proficiency standards. We conclude that the heterogeneity among the states' NAEP equivalents is mainly due to the variation in standards established by the states. Again, that variation arises not simply from the

positions of the standards on the state test score scales but from the totality of factors that determine student performance on the state tests and on NAEP.

Clearly, one can posit different scenarios that offer alternative explanations for the wide range in percent proficient that has been observed. What might be one such scenario? Suppose that two states (denoted A and B) employ the same test for accountability, which differs from NAEP in the relative emphasis placed on the different content strands. In particular, imagine that there is one strand that is strongly represented on the state test but hardly at all on the NAEP assessment. Suppose further that the states set their proficiency thresholds at the same point on the scale. Thus, by construction, their standards are of equal stringency. Now if the students in State A are better prepared for the state test (with special attention to that one strand) than students in State B, then the distribution of scores in State A will be stochastically larger than that in State B and, perforce, the percent proficient in State A will be greater than the percent proficient in State B. However, State A's advantage is not reflected in the NAEP distributions of the two states. Consequently, the NAEP equivalent for State A will be lower than that for State B—and one would conclude (incorrectly) that State A's proficiency standard is less stringent than State B's.

Could an approximation to such a scenario, aggregated over a number of pairs of states, have plausibly generated Figures 12 and 13? We argue in the negative. First, because assessment frameworks do not differ substantially in, say, Grade 4 mathematics. Consequently, differences in emphasis are not likely to lead to substantial differences in percent proficient that are not accompanied by corresponding differences in NAEP distributions. That is, observing the range in the percent proficient similar to that in Figures 12 and 13 is implausible under this scenario. Moreover, under this scenario, if it were the case that states with the higher values of the percent proficient were being penalized by the linking method for their superior performance on the state tests that is not reflected in NAEP, then one might expect that those states would display lower within-state correlations between an indicator of state test performance and NAEP scores. We carried out this computation for the states in Figures 12 and 13, after dividing the states into two groups based on a median split on the percent proficient. For each state, we calculated the Spearman correlation across schools between the percent proficient on the state test and the estimated mean on NAEP. The mean correlations in the two groups were nearly identical. [12]

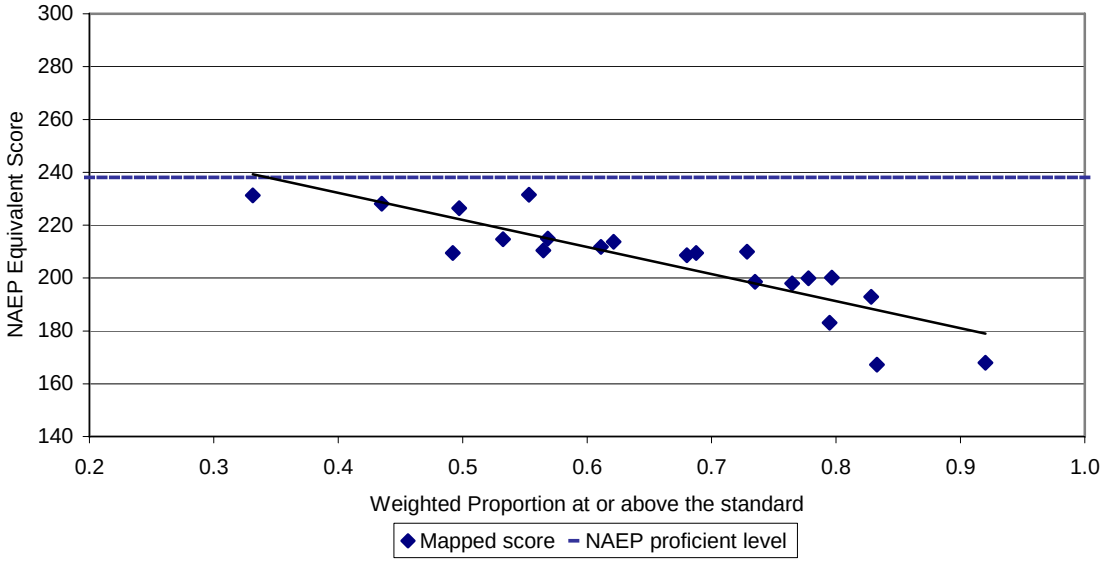

Figure 22. **G4 2002 reading: NAEP equivalents to the state standards of proficient vs. proportions at or above state standards of proficient.**

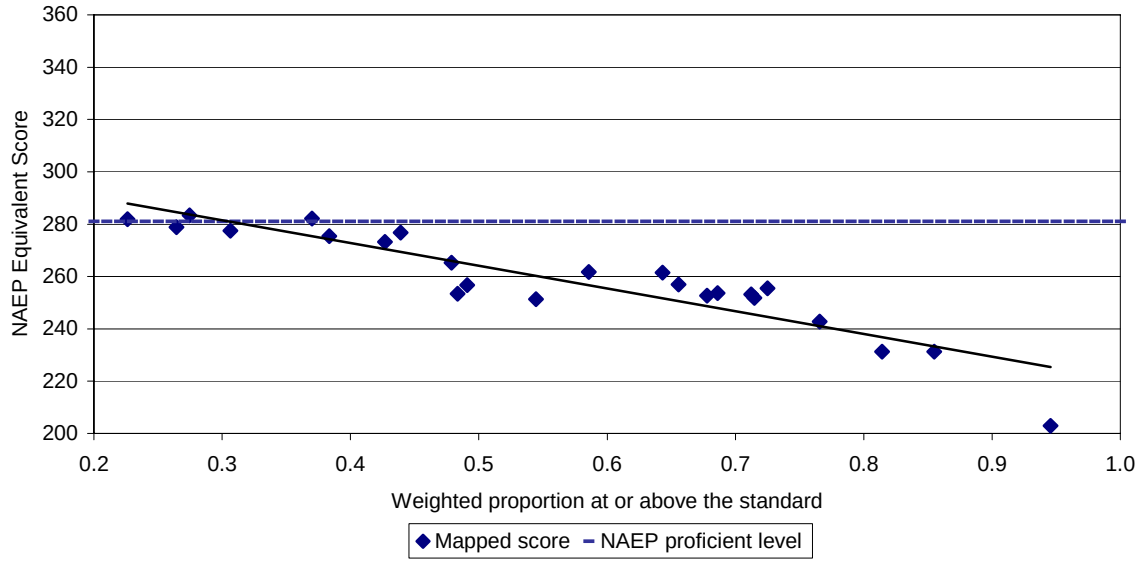

Figure 23. **G8 2002 reading: NAEP equivalents to the state standards of proficient vs. proportions at or above state standards of proficient.**

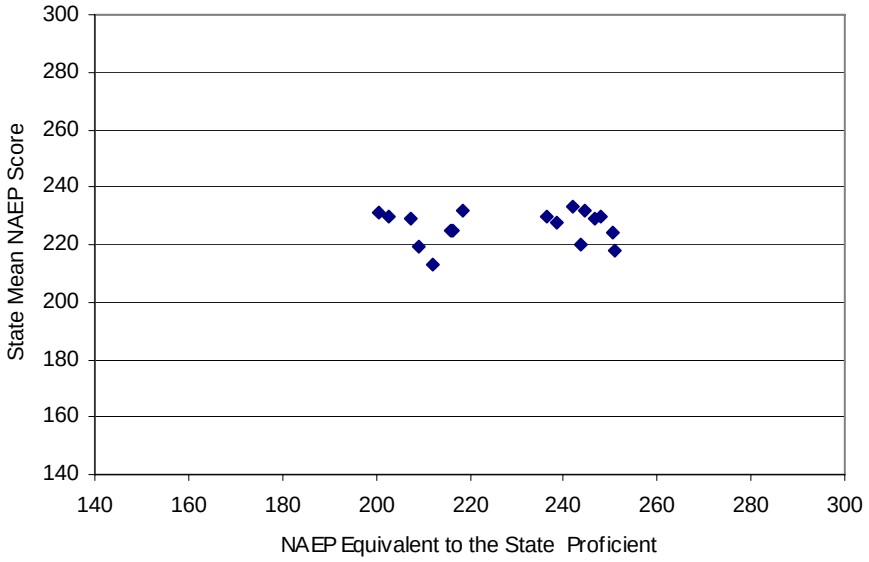

Figure 24. **G4 2000 math: NAEP equivalent scores to state proficient standards vs. state mean NAEP scores.**

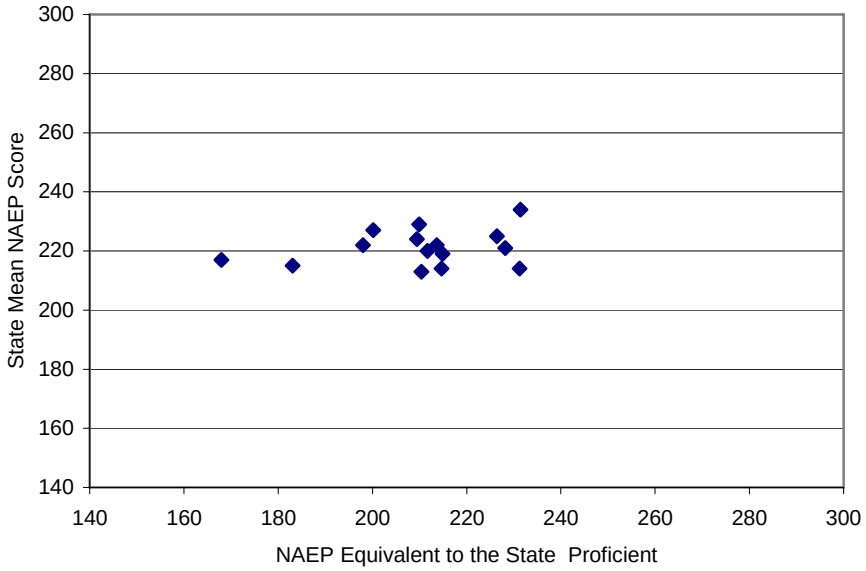

Figure 25. **G4 2002 reading: NAEP equivalent scores to state proficient standards vs. state mean NAEP scores.**

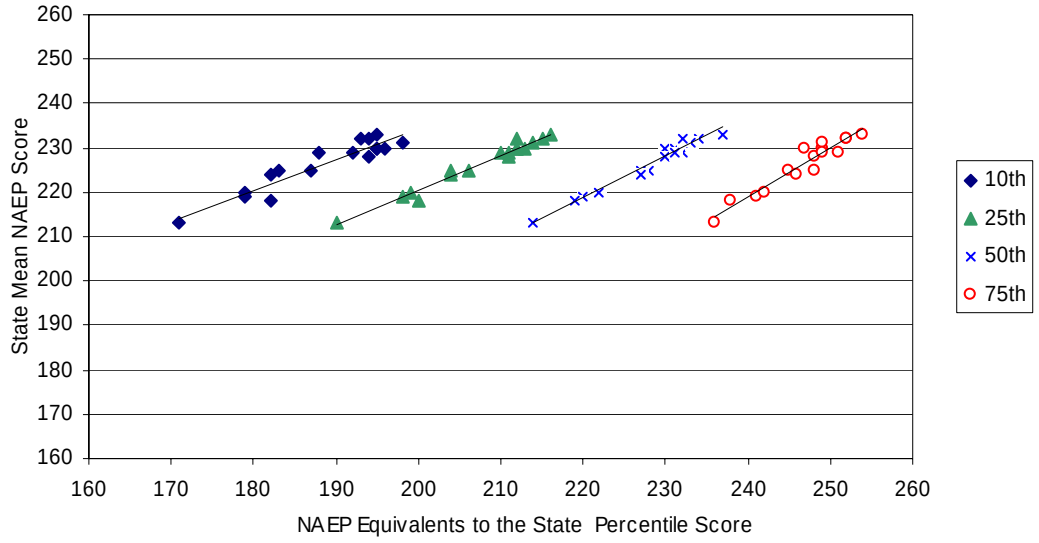

Figure 26. **G4 2000 math: State mean NAEP scores vs. NAEP equivalents to the state percentile scores.**

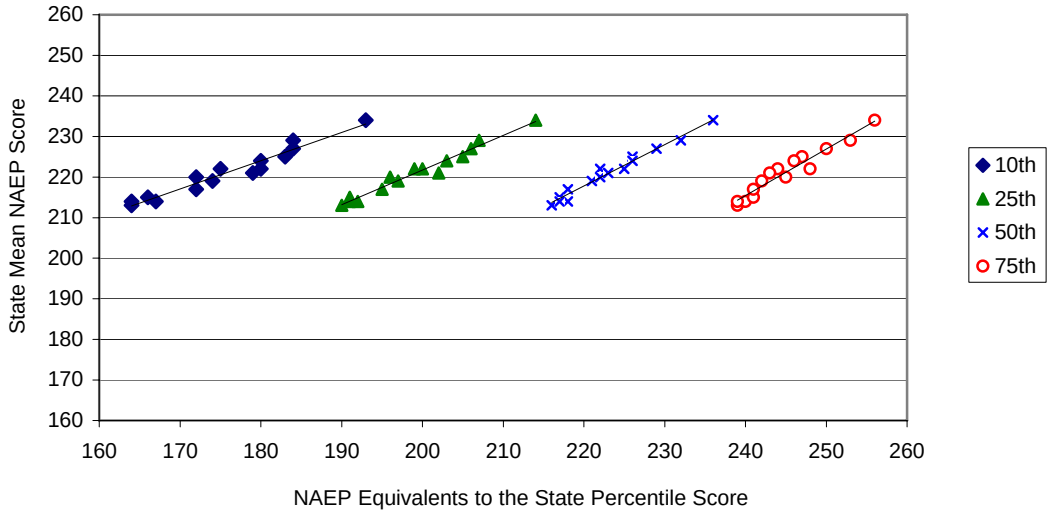

Figure 27. **G4 2002 reading: State mean NAEP scores vs. NAEP equivalents to the state percentile scores.**

Another scenario focuses specifically on differences in curriculum. Suppose, for example, there is a state in which a substantial proportion of the math curriculum content for the 4th grade reflects concepts and procedures that are covered at the third grade level in most other states and, as a result, there is incomplete coverage of typical 4th grade content. Suppose further that the state's test is a valid test of the state's curriculum and that the standard is set at a moderately high level. Nonetheless, students taking the NAEP assessment will encounter many questions for which they are generally not prepared, with the result that the state's NAEP distribution will likely be stochastically smaller than those for states with curricula that are better matched to the NAEP framework. The state's NAEP equivalent score will tend to fall at the low end of the range. This example highlights the point that a state's equivalent score may well reflect more than just the placement of the proficiency standard. However, since the range of the state NAEP means is only about 20 points, the difference among the states in student preparation is unlikely to be the main factor in explaining the results depicted in Figures 12 and 13.

That state standards for proficiency can apparently differ by 50 or more points on the NAEP scale should give pause both to policy makers and educators. What, indeed, is expected of students in states with the lowest NAEP equivalents? How do these expectations differ from states with the highest NAEP equivalents? What does the achievement of proficiency signify in terms of what students know and can do? In our view, mapping state standards to the NAEP scale makes possible conversations that could be more constructive than simple comparisons of percent above standard. A relative low NAEP equivalent is a warning signal to the state that what it expects of its students may differ materially from the expectations in other states. In particular, it should provide greater impetus to carry out an intensive cross-state analysis of content and performance standards.

After consideration of a number of scenarios, we believe we are on safe ground with the assertion that our results support the contention that differences across states in performance expectations, as manifested in the apparent stringency of the proficiency standards, remain the most plausible explanation of the heterogeneity in percent proficient. At the same time, we recognize that the issue cannot be settled directly unless states adopt a common content framework and implement a common examination based on that framework. Because that is unlikely, we must accommodate to the inherent ambiguity in the situation. Thus, we should

certainly refrain from making fine distinctions among NAEP equivalents. At the very least, confidence bands, based on the estimated standard errors, should be used for all comparisons, with the recognition that they do not capture all of the uncertainty that attaches to the NAEP equivalents for the intended inferences.

6. Another Application: Mapping the NAEP Achievement Standards Onto a State Test Scale

When state standards are mapped onto the NAEP scale, we can compare and evaluate the different standards despite the differences in tests and standard-setting procedures. The application described in this section is a reverse mapping procedure; that is, finding a point on the state test score scale that best corresponds to the NAEP achievement cut point. These state equivalents to the NAEP achievement levels could provide state educators and policy makers with useful information to directly compare their standards to national benchmarks.

Figure 28 illustrates the reverse mapping procedure, which, as before, is based on the principle of equipercentile equating. Although the figure is analogous to that of Figure 1, the direction of the mapping is reversed: going from right to left. The curve on the right side represents the estimated distribution of NAEP scores for the students sampled in the state. The point on the NAEP scale is the cut point of a NAEP achievement level, which represents one of the NAEP standards: *basic*, *proficient*, or *advanced*. Let the upper tail area be equal to $\hat{p}_w$. The curve on the left side represents the distribution of scores on the state test of all students in all schools in the state. The estimated state equivalent standard of the NAEP achievement level is the point on the state scale above which the tail area is also equal to $\hat{p}_w$.

To accomplish the reverse mapping, the actual distribution of state test scores is required. (That is why the distribution is represented by a solid line, rather than a dashed line as in Figure 1.) Unfortunately, actual student scores for most states are not contained in the NLSLSASD database. Accordingly, we were only able to conduct a case study for the Michigan G4 2000 state mathematics test, for which the appropriate data were available. [13]

The reverse mapping procedure also employs the jackknife replicate resampling (JRR) approach to estimate the variances for the sampling and measurement errors, as described in Section 4. The procedure uses the distribution of student scores to calculate $\hat{p}_w$, rather than the proportions of students in each school meeting the standard. Therefore, the reverse procedure

employs student design weights to estimate the distribution, and the replicate weights for the JRR procedure are also computed from student design weights. Again, measurement error is estimated from repeating the procedure for each set of plausible values.

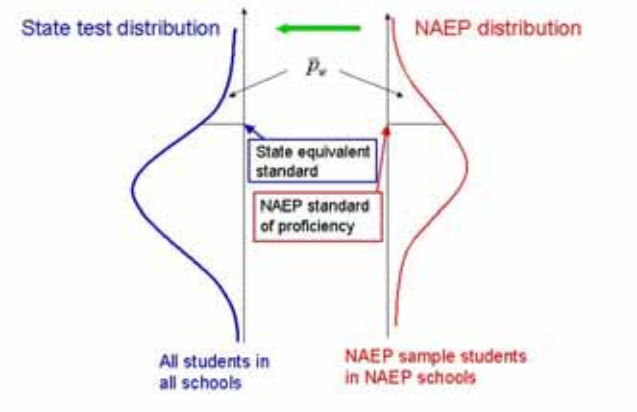

Figure 28. **Schematic for the reverse mapping.**

Table 9 presents the state equivalents to the NAEP mathematics achievement levels and their standard errors. The mapped NAEP achievement levels on the Michigan state test scale are 518, 554, and 595 for *basic*, *proficient*, and *advanced* levels, [14] respectively. The corresponding percentages of students meeting these levels are about 70.2, 27.5, and 2.8, respectively.

The Michigan state test score distribution indicates that the percentages of students meeting state standards, *moderate* and *satisfactory*, are 91.3 and 75.1, respectively. It appears that the standard of *satisfactory* is set at a level lower than the *basic* level of the NAEP mathematics achievement.

7. Conclusions and Recommendations

The purpose of this study was to continue methodological development of an approach originally proposed by McLaughlin and associates for making useful comparisons among state standards. (We again emphasize that this mapping procedure should NOT be used to make high-stakes decisions about schools or districts.) It is assumed that the state assessment and the NAEP assessment reflect similar content and have comparable structures, although they differ in test and item formats as well as standard-setting procedures. This development consisted of two

modifications: (a) a shift from a school-based to a student-based strategy for estimating the NAEP equivalent to a state standard, and (b) the derivation of a more refined estimate of the variance of the NAEP equivalent by taking into account the NAEP design in the calculation of sampling error and by obtaining an estimate of the contribution of measurement error.

The new methodology was applied to four sets of data: (a) year 2000 state mathematics tests and the NAEP 2000 mathematics assessments for Grades 4 and 8, and (b) year 2002 state reading tests and the NAEP 2002 reading assessments for Grades 4 and 8. For the first dataset, we also applied the method described by McLaughlin and associates. We found that for both mathematics and reading, there is a strong negative linear relationship across states between the proportions meeting the standard and the apparent stringency of the standard as indicated by its NAEP equivalent.

Table 9

The State Equivalents to the NAEP Mathematics Achievement Levels and Their Standard Errors for 2000 Michigan State Mathematics Test, Grade 4

	Basic	Proficient	Advanced
NAEP achievement level	214	249	282
State equivalent standard	518	554	595
SE due to sampling error	1.21	3.14	0.98[a]
SE due to measurement error	0.79	0.30	0.14
Total SE	1.45	3.16	0.99

[a] On average, only 2.8% of Michigan students meet the mapped standard for *advanced*. Therefore, number of students at each school meeting the standard is small and results in a jackknifed variance that is very large, 51.39. In particular, the 41st replicate contributes about 98% of the total sampling variation. Evidently, this is a very problematic estimate. After considering several approaches, we decided to use the Winsorized variance estimate, shown in this table. In the calculation of the Winsorized estimate, the largest and smallest of the squared deviations are replaced by their nearest neighbor values.

Comparable results can be found in a recent report by Kingsbury, Olson, Cronin, Hauser, and Houser (2003) describing an effort to map the proficiency standards for 12 states onto a common scale, which is used to report test scores for the Northwest Evaluation Association (NWEA) assessment battery. This exercise was carried out in both reading and mathematics for Grades 3–10, employing data collected between 1999 and 2003. In contrast to the present case, NWEA has available individual student scores on both the state test and the (common) NWEA scale. The authors also found substantial heterogeneity among the NWEA equivalents of the state proficiency standards as well as a strong negative correlation between the percent proficient and the NWEA equivalent to the state's proficiency standard. Although the NWEA linking methods as well as the data have both strengths and weaknesses in comparison to the exercise described in this chapter, it is instructive to compare the results of the two approaches. We did so for 2000 mathematics in Grades 4 and 8 and for 2002 reading in Grades 4 and 8.[15] There is good agreement between the rankings of the states on the apparent stringency of their proficiency standards, adding to the credibility of our findings.

Recall that the motivation for attempting to map state standards onto a common scale was to account for the observed differences among states in the proportions of students declared proficient. The credibility and utility of the approach depends on making two arguments: first, that the estimated NAEP equivalents are both well estimated and stable; second, that one can attribute the differences in NAEP equivalents across states to differences in performance standards and, in some cases, to differences in content standards as well. If the two arguments are established, then the results obtained herein indicate that the most important factor in explaining why two states have substantially different proportions of students meeting their respective proficiency standards is the comparative rigor of both their performance standards and content standards.

With respect to the first argument, the estimated standard deviations of the NAEP equivalents, taking into account both sampling and measurement errors, are generally small in comparison to the range of the NAEP equivalents. Stability is best addressed by implementing the linkage for different subgroups. As we have already indicated, that is possible only for a few states. An alternative is to examine, for each state, the correlation between performance on the state test and on NAEP. This can be done at the school level. For example, using the NLSLSASD files, for each state one can compute the raw Spearman correlation across schools

between the percent proficient on the state test and the estimated NAEP mean. For Grade 4 mathematics, the median correlation is about .7. Ideally, one would like to supplement the quantitative analysis with an intensive examination of the degree of alignment between the state test frameworks and the NAEP frameworks. This has not been done.

With respect to the second argument, the essential difficulty is that one must reason from the observed results (e.g., Figures 12 and 13) back to the true state of nature. The plausibility of the second argument is supported by the observation that there is a weak relationship between states' percent proficient and states' performance on NAEP. There is also a weak relationship between states' NAEP means and their NAEP score equivalents. Note also that the heterogeneity among the NAEP equivalents is much greater than among NAEP means. It is possible to construct alternative scenarios that are consistent with Figures 12 and 13 but lead to different inferences about the relative stringency of state standards. However, as discussed in Section 5.3, the evidence in Figures 26 and 27 supports the contention that the heterogeneity among the NAEP equivalents largely reflects differences in the rigor of their proficiency standards. Indeed, if states were only to set their proficiency standards at a common fixed percentile, we would have observed a much more credible relationship between NAEP equivalents and NAEP means (e.g., Figure 26) rather than what was actually observed (e.g., Figure 24) and this despite the considerable differences among states in their assessment programs.

In view of the limitations of the data available, inferences concerning the NAEP score equivalents should be made with due caution. As indicated at the outset, in some states a number of schools in the NAEP sample could not be included in the analysis, because the required state test data were not available at the individual school level. The loss of these schools could introduce some bias. In other states, the relevant state assessment was labeled *English/Language Arts* rather than *Reading*, so the degree of alignment between the two assessments could be lower than for other states. In any case, for each subject and grade combination, state assessment frameworks, as well as the test structures and item formats employed, will differ from those of the corresponding NAEP assessment. These differences can add noise to the comparisons with NAEP. Additionally, states differ in the numbers and proportions of students with disabilities or English language learners that are excluded from either the state assessment or NAEP (or both). Such differences can also contribute to differences in the estimated NAEP score equivalents. Consequently, the estimated variance associated with each NAEP equivalent provides only a

lower limit to the uncertainty to be associated with that value. At the same time, it is highly unlikely that the sources of bias discussed above could yield the broad range of NAEP score equivalents obtained.

Finally, we note that under NCLB, a state's NAEP results are to be used to confirm its success in achieving adequate yearly progress. Currently, such a confirmation is based on observing changes at the mean of the distribution of the state test and changes at the mean of the state's NAEP distribution. It is possible to use changes in the estimated NAEP equivalent over time in a similar manner. For example, if the proportion above the proficient standard on a state's test increases over time while the NAEP distribution remains constant, then the estimated NAEP equivalent would correspondingly decrease. It is possible, but not obvious, that tracking changes in the NAEP equivalent is to be preferred to tracking changes in the mean for the purpose of monitoring state outcomes. At the same time, interpreting trends in state test scores is problematic in view of the many factors that can impact score levels. Attempting to do so in terms of linkages to another test (e.g., NAEP) is more problematic still, because of the many ways in which the invariance of the linkage over time might fail. This is likely to be the case no matter which feature of the distributions is selected. For more on these issues, consult Thissen (2007) and Koretz (2007).

References

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

Braswell, J., Lutkus, A., Grigg, W., Santapau, S., Tay-Lim, B., & Johnson, M. (2001). *The nation's report card: Mathematics 2000*. Washington DC: National Center for Education Statistics.

Braun, H. I., & Holland, P. W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D.B. Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic Press.

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

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.

Grigg, W., Daane, M., Jin, Y., & Campbell, J. (2003). *The nation's report card: Reading 2002*. Washington DC: National Center for Education Statistics.

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

Kingsbury, G. G., Olson, A., Cronin, J., Hauser, C., & Houser, R. (2003). *The state of state standards: Research investigating proficiency levels in fourteen states*. Portland OR: Northwest Evaluation Association.

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

Koretz, D., 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.

Koretz, D. (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. 229-353). New York: Springer-Verlag.

Lane, S. (2004). 2004 NCME presidential address: Validity of high-stakes assessments: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice*, 23(3), 6-14.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education* 6(1), 83-102.

Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives 11(31)*. Retrieved January 20, 2004, from http://epaa.asu.edu/epaa/v11n31/

McLaughlin, D. (2000). *Protecting state NAEP trends from changes in SD/LEP inclusion rates.* Palo Alto, CA: American Institutes for Research.

McLaughlin, D., & Bandeira de Mello, V. (2002, April). *Comparison of state elementary school mathematics achievement standards, using NAEP 2000*. Paper presented at the American Educational Research Association Annual Meeting, New Orleans, LA.

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.

Pitoniak, M. J., & Mead, N. A. (2003). *Statistical methods to account for excluded students in NAEP: 2002 reading and writing assessments*. Princeton, NJ: ETS.

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. 193-226). Washington, DC: National Center for Education Statistics.

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-Verlag.

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

Notes

[1] This research was carried out while Henry Braun was a distinguished presidential appointee at ETS.

[2] For a general introduction to NAEP, see Jones & Olkin (2004).

[2] Data from Grades 4 and 8 were analyzed in this report.

[4] 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 NCES. Its purpose is to collect and validate data from state testing programs across the country. It contains assessment data for approximately 80,000 public schools in the United States and is updated annually.

[5] This result suggests that using *P* instead of should yield similar results. This point is addressed in Section 7.

[6] For reporting purposes, two sample types were formed in the operational NAEP assessment: R2 and R3. The sample type R2 provides inferences for a less inclusive population where accommodations were not permitted; the sample type R3 provides inferences for a more inclusive population where accommodations were permitted.

[7] The estimates for Maine, Montana, Louisiana, and North Carolina can be found in Tables 2 and 6.

[8] The estimates for Arizona and Louisiana can be found in Tables 2 and 6.

[9] Some of the state standards for proficiency were selected by their names and others were inferred by the authors. The standards so designated are marked by asterisks in the first column of Tables 5 and 6 and Tables 7 and 8.

[10] While NAEP mathematics for Grades 4 and 8 was jointly scaled in 1990, the cross-grade property has not been retained in order to focus on within-grade trends over time. Accordingly, between-grade comparisons for the 2000 administration cannot be formally supported.

[11] For Grade 4 mathematics, the coefficient of variation of the NAEP equivalents is about 19 times larger than that for the NAEP means. For Grade 4 reading the ratio is about 9. For Grade 8 mathematics the ratio is about 18 and for Grade 8 reading, it is about 16.

[12] The Spearman correlations for two groups are .69 and .73 separately. For this calculation, Nebraska was set aside as an outlier.

[13] The 2000-2001 Michigan student level data were publicly available (http://www.schooldata.org).

[14] The three cut points of NAEP achievement levels, *basic*, *proficient*, and *advanced*, are 214, 249 and 282, respectively (Braswell et al., 2001).

[15] Unfortunately the overlap among states for which data are available is not as great as one would hope, being greater in Grade 8 than in Grade 4.

Appendix

While the results obtained by Braun and Qian (WAM) are qualitatively similar to those obtained by McLaughlin and associates (ULM), there are both conceptual and technical differences that are worth noting. Not surprisingly, we conclude that those differences favor the approach we adopted for the study. We summarize the argument below.

1. An important source of the difference in the results between WAM and ULM is that we are each actually estimating different quantities. For WAM, the target is

$$y_{WAM} = G^{-1}(1-P)$$

where

P = proportion of students in the state meeting the standard defined on the state test scale.

G = distribution for students in the state on the NAEP scale.

Note that the target y_{WAM} is defined for a population of students.

For ULM, the target is

$$\bar{z}_{ULM} = \underset{k}{\text{ave}}\left\{G_k^{-1}(1-P_k)\right\}$$

where

P_k = proportion of students in school k meeting the standard defined on the state test scale.

G_k = distribution for students in school k on the NAEP scale.

and the (simple) average is taken over all schools in the state. Note that the target $\bar{z}_{ULM}$ is defined for a population of schools.

2. The ULM results claim that

$$Z_k = G_k^{-1}\left(1 - P_k\right)$$

are all equal. We don't believe this can be true, given observed correlations between state test scores and NAEP scores and the differences between schools. Rather, we take the intuitive argument put forward by McLaughlin and associates to indicate that, while the P_k may vary substantially among schools, the variation among Z_k is considerably smaller. (Were this not the case, it is not clear that the target $\overline{z}_{ULM}$ would be very meaningful. But see point 5b below.)

3. In general, y_{WAM} and $\overline{z}_{ULM}$ are not equal. To see this, let us consider two special cases. (For the moment, we assume all students take both tests.)

   a. Suppose there are *K* schools of equal size. Then

$$y_{WAM} = G^{-1}\left(\underset{k}{\text{ave}}\left\{1 - P_k\right\}\right)$$

   and

$$\overline{z}_{ULM} = \underset{k}{\text{ave}}\left\{G_k^{-1}\left(1 - P_k\right)\right\}.$$

   But

$$G = \underset{k}{\text{ave}}\left\{G_k\right\},$$

   so $y_{WAM} = \overline{z}_{ULM}$ would imply

$$G^{-1}\left(\underset{k}{\text{ave}}\left(1 - P_k\right)\right) = \left[\underset{k}{\text{ave}}\left\{G_k\right\}\right]^{-1}\left(\underset{k}{\text{ave}}\left(1 - P_k\right)\right) = \underset{k}{\text{ave}}\left\{G_k^{-1}\left(1 - P_k\right)\right\}.$$

We know of no theorem that would assert this, even in the case that all the $G_k^{-1}(1-P_k)$ were equal to a common value, $\bar{z}_{ULM}$.

b. Suppose there are only two schools in the state: S_1 with 100 students and S_2 with 1000 students. Suppose further that $Z_1 = 250$ and $Z_2 = 260$. Then

$$z = (250 + 260)/2 = 255.$$

But

$$y_{WAM} > 255,$$

since the estimates of *P* and *G* will be dominated by the data from S_2. Of course, this can be made more realistic by increasing the number of schools, but the point is the same.

4. If the two approaches are indeed estimating different targets, which one is to be preferred? Not surprisingly, we assert that the target of WAM is the more appropriate. First, the between-state comparisons that sparked this effort have been usually framed in terms of the proportions of students meeting a standard, rather than the average of school proportions meeting the standard.

   Second, consider two states with the same distributions of student scores on a common test, as well as on NAEP, but with different allocations of students to schools. Our argument in point 3 above shows that the ULM method would yield different estimates, while the WAM method would necessarily yield the same result.

5a. From the outset, we were concerned that ULM does not employ weights in its estimation procedures. We agree that the task of estimating standards is different from that of estimating a population mean. In the latter case, sampling weights must be used. In the former case, they are less critical but their neglect could still be problematic. Differences among schools in estimated standards may be real (i.e., not

just due to measurement error), in which case a weighted average may be a more desirable target than an unweighted one. This would certainly be the case if there were a correlation between school size and the proportion of students meeting the standard.

5b. To this end, we plotted $\tilde{z}_k$ against school size separately for the three California standards of 2000 mathematics, where the $\tilde{z}_k$ are obtained by applying D. McLaughlin's method to the NAEP reported data (i.e., not to the FPE data). Figures A1-A2 for *CA(25)*, *CA(50)*, and *CA(75)* present the results. Recall that the $\tilde{z}_k$'s are on the NAEP scale. In each case, there is a statistically significant linear regression. Moreover, for each standard there is a broad range for the $\tilde{z}_k$ values, along with a substantial overlap across standards. For example, more than half of the $\tilde{z}_k$'s for *CA(25)* fall in the range of the $\tilde{z}_k$'s for *CA(75)*. [Note that we don't have access to the z_k employed by D. McLaughlin, so that we cannot produce analogous figures for them.]

5c. We also have concerns about the failure of ULM to take account of the finite population correction (FPC), as well as the uncertainty due to measurement error, into the variance estimate. If the sample at hand constituted a census of schools in the state, then the variance formula employed for ULM could not represent sampling variance. Rather it would reflect the heterogeneity in the Z_k among schools (refer to 5b above). This is a quantity of some interest, but it is not what we are after. Similarly, the Z_k are obtained by evaluating $\hat{G}_k^{-1}(\cdot)$ at the point $(1-P_k)$, where $\hat{G}_k^{-1}(\cdot)$ is an estimate of the NAEP distribution for the school. This estimate can be quite variable and to ignore this uncertainty in calculating the sampling variance of $\overline{z}_{ULM}$ seems unwarranted.

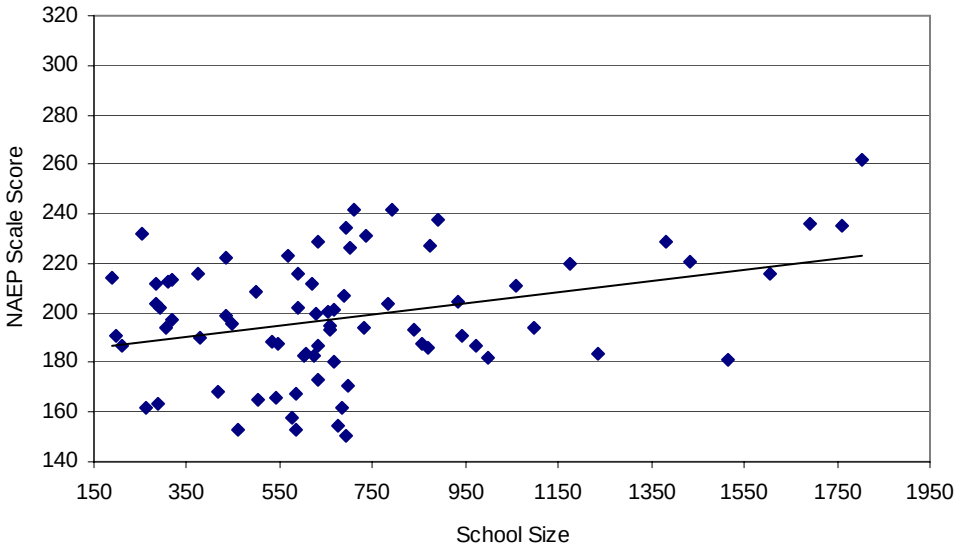

Figure A1. **NAEP scale score vs. school size for California G4 2000 math, *CA(25).***

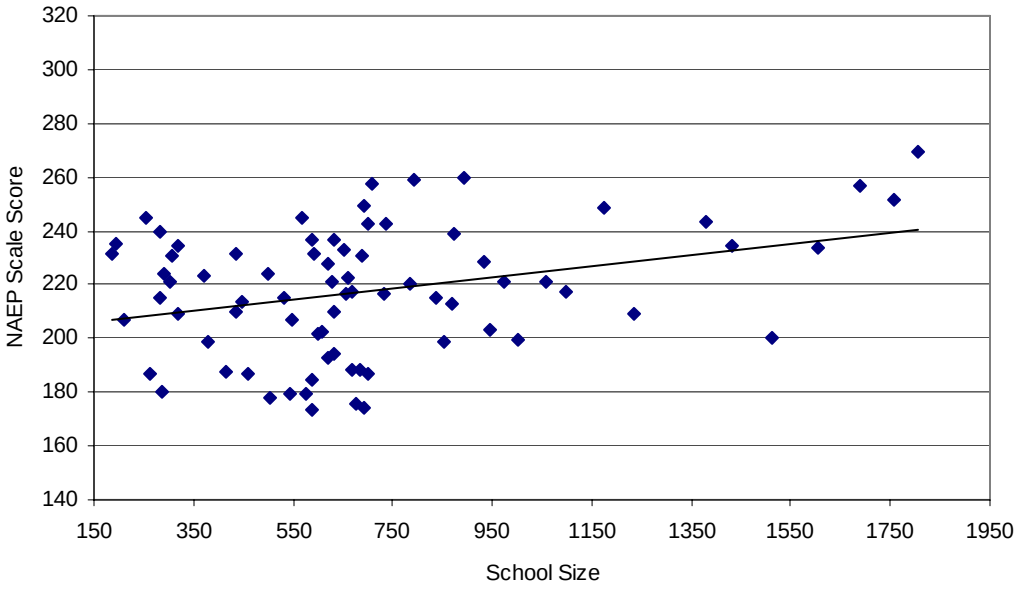

Figure A2. **NAEP scale score vs. school size for California G4 2000 Math, *CA(50).***