Mapping State Proficiency Standards Onto NAEP Scales: 2005-2007

NCES 2010-456
U. S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

U.S. DEPARTMENT OF EDUCATION
NCES 2010-456

Mapping State Proficiency Standards Onto NAEP Scales: 2005-2007

Research and Development Report

October 2009

Victor Bandeira de Mello
Charles Blankenship
American Institutes for Research

Don McLaughlin
Statistics and Strategies

Taslima Rahman
Project Officer
National Center for Education Statistics

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Statistics
Stuart Kerachsky
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high-priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high-quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public. Unless specifically noted, all information contained herein is in the public domain.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

October 2009

The NCES World Wide Web Home Page address is http://nces.ed.gov.
The NCES World Wide Web Electronic Catalog address is http://nces.ed.gov/pubsearch.

Suggested Citation
Bandeira de Mello, V., Blankenship, C., and McLaughlin, D.H. (2009). *Mapping State Proficiency Standards Onto NAEP Scales: 2005-2007* (NCES 2010-456). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

For ordering information on this report, write to

U.S. Department of Education
ED Pubs
P.O. Box 1398
Jessup, MD 20794-1398

or call toll free 1-877-4ED-Pubs or order online at http://www.edpubs.org.

Content Contact
Taslima Rahman
(202) 502-7316
taslima.rahman@ed.gov

FOREWORD

The Research and Development (R&D) series of reports at the National Center for Education Statistics has been initiated to

- Share studies and research that are developmental in nature. The results of such studies may be revised as the work continues and additional data become available;
- Share the results of studies that are, to some extent, on the cutting edge of methodological developments. Emerging analytical approaches and new computer software development often permit new and sometimes controversial analyses to be done. By participating in frontier research, we hope to contribute to the resolution of issues and improved analysis; and
- Participate in discussions of emerging issues of interest to educational researchers, statisticians, and the federal statistical community in general.

The common theme in all three goals is that these reports present results or discussions that do not reach definitive conclusions at this point in time, either because the data are tentative, the methodology is new and developing, or the topic is one on which there are divergent views. Therefore, the techniques and inferences made from the data are tentative and subject to revision. To facilitate the process of closure on the issues, we invite comment, criticism, and alternatives to what we have done.

Such responses should be directed to

Marilyn Seastrom
Chief Statistician
Statistical Standards Program
National Center for Education Statistics
1990 K Street NW
Washington, DC 20006-5651

EXECUTIVE SUMMARY

Since 2003, the National Center for Education Statistics (NCES) has sponsored the development of a method for mapping each state's standard for proficient performance onto a common scale—the achievement scale of the National Assessment of Educational Progress (NAEP). When states' standards are placed onto the NAEP reading or mathematics scales, the level of achievement required for proficient performance in one state can then be compared with the level of achievement required in another state. This allows one to compare the standards for proficiency across states.

The mapping procedure offers an approximate way to assess the relative rigor of the states' adequate yearly progress (AYP) standards established under the No Child Left Behind Act of 2001. Once mapped, the NAEP scale equivalent score representing the state's proficiency standards can be compared to indicate the relative rigor of those standards. The term rigor as used here does not imply a judgment about state standards. Rather, it is intended to be descriptive of state-to-state variation in the location of the state standards on a common metric.

This report presents mapping results using the 2005 and 2007 NAEP assessments in mathematics and reading for grades 4 and 8. The analyses conducted for this study addressed the following questions:

- How do states' 2007 standards for proficient performance compare with each other when mapped on the NAEP scale?
- How do the 2007 NAEP scale equivalents for state standards compare with those estimated for 2005?
- Using the 2005 NAEP scale equivalent for state standards to define a state's *proficient* level of performance on NAEP, do NAEP and that state's assessment agree on the changes in the proportion of students meeting that state's standard for proficiency from 2005 to 2007?

To address the first question, the 2007 *NAEP scale equivalent* of each state reading and mathematics proficiency standard for each grade was identified. The mapping procedure was applied to the test data of 48 states.[1] Key findings of the analysis presented in Section 3 of the report are:

- In 2007, as in 2003 and 2005, state standards for *proficient* performance in reading and mathematics (as measured on the NAEP scale) vary across states in terms of the levels of achievement required. For example, the *distance* separating the five states with the highest standards and the five states with the lowest standards in grade 4 reading was comparable to the difference between *Basic* and *Proficient* performance on NAEP.[2] The distance was as large in reading at grade 8 and as large in mathematics in both grades.

[1] Test data for the District of Columbia, Nebraska, and Utah were not available to be included in the analysis. California does not test general mathematics in grade 8.

[2] NAEP defines *Proficient* as *competency over challenging subject matter*, not grade-level performance. *Basic* is defined as partial mastery of the skills necessary for *Proficient* performance.

- In both reading and mathematics, the 29- to 30-point distance separating the five highest and the five lowest NAEP scale equivalent of state standards for *proficient* performance was nearly as large as the 35 points that represent approximately one standard deviation in student achievement on the NAEP scale.

- In grade 4 reading, 31 states set grade 4 standards for proficiency (as measured on the NAEP scale) that were lower than the cut point for *Basic* performance on NAEP (208). In grade 8 reading, 15 states set standards that were lower than the *Basic* performance on NAEP (243).

- In grade 4 mathematics, seven states set standards for proficiency (as measured on the NAEP scale) that were lower than the *Basic* performance on NAEP (214). In grade 8 mathematics, eight states set standards that were lower than the *Basic* performance on NAEP (262).

- Most of the variation (approximately 70 percent) from state to state in the percentage of students scoring proficient or above on state tests can be explained by the variation in the level of difficulty of state standards for proficient performance. States with higher standards (as measured on the NAEP scale) had *fewer* students scoring proficient on state tests.

- The rigor of the state standards is not consistently associated with higher performance on NAEP. This association is measured by the squared correlation between the NAEP scale equivalent of the state standards and the percentages of students who scored at or above the NAEP *Proficient* level. In grade 4 reading and mathematics, the squared correlations are around .10 and statistically significant. In grade 8 reading and mathematics, the squared correlations are less than .07 and are not statistically significant.

To address the second question, the analyses focused on the consistency of mapping outcomes over time using both 2005 and 2007 assessments. Although NAEP did not change between 2005 and 2007, some states made changes in their state assessments in the same period, changes substantial enough that states indicated that their 2005 scores were not comparable to their 2007 scores. Other states indicated that their scores for those years are comparable. Comparisons between the 2005 and 2007 mappings in reading and mathematics at grades 4 and 8 were made separately for states that made changes in their testing systems and for those that made no such changes.[3] Key findings of the analysis presented in Section 4 are:

- In grade 4 reading, 12 of the 34 states with available data in both years indicated substantive changes in their assessments. Of those, eight showed significant differences between the 2005 and 2007 estimates of the NAEP scale equivalent of their state standards, half of which showed an increase and half a decrease.

- In grade 8 reading, 14 of the 38 states with available data in both years indicated substantive changes in their assessments. Of those, seven showed significant differences between the 2005 and 2007 estimates of the NAEP scale equivalent of their state standards, all seven showed lower 2007 estimates of the NAEP scale equivalents.

[3] The 2005 mappings in this report will not necessarily match previously published results (U.S. Department of Education 2007). Methodological differences between the procedures used in both analyses will generally cause empirical results to show small differences that are not large enough to change the whole-number scale value reported as the NAEP equivalent.

- In grade 4 mathematics, 14 of the 35 states with available data in both years indicated substantive changes in their assessments. Of those, 11 showed significant differences between the 2005 and 2007 estimates of the NAEP scale equivalent of their state standards: 6 states showed a decrease and 5 showed an increase.

- In grade 8 mathematics, 18 of the 39 states with available data in both years indicated substantive changes in their assessments. Of those, 12 showed significant differences between the 2005 and 2007 estimates of the NAEP scale equivalent of their state standards: 9 showed a decrease and 3 showed an increase.

For the states with no substantive changes in their state assessments in the same period, the analyses presented in Section 4 indicate that for the majority of states in the comparison sample (14 of 22 in grade 4 reading, 13 of 24 in grade 8 reading, 15 of 21 in grade 4 mathematics and 14 of 21 in grade 8 mathematics), the differences in the estimates of NAEP scale equivalents of their state standards were not statistically significant.

To address the third question, NAEP and state changes in achievement from 2005 to 2007 were compared. The percentage of students reported to be meeting the state standard in 2007 is compared with the percentage of the NAEP students in 2007 that is above the NAEP scale equivalent of the same state standard in 2005. The analysis was limited to states with (a) available data in both years and (b) no substantive changes in their state tests. The number of states included in the analyses ranged from 21 to 24, depending on the subject and grade. The expectation was that both the state assessments and NAEP would show the same changes in achievement between the two years. Statistically significant differences between NAEP and state measures of changes in achievement indicate that more progress is made on either the NAEP skill domain or the state-specific skill domain between 2005 and 2007. A more positive change on the state test indicates students gained more on the state-specific skill domain. For example, a focus in instruction on state-specific content might lead a state assessment to show more progress in achievement than NAEP. Similarly, a less positive change on the state test indicates students gained more on the NAEP skill domain. For example, focus in instruction on NAEP content that is not a part of the state assessment might lead the state assessment to show progress in achievement that is less than that of NAEP. Key findings from Section 5 are:[4]

- In grade 4 reading, 11 of 22 states showed no statistically significant difference between NAEP and state assessment measures of changes in achievement; 5 states showed changes that are more positive than the changes measured by NAEP, and 6 states showed changes that are less positive than those measured by NAEP.

- In grade 8 reading, 9 of 24 states showed no statistically significant difference between NAEP and state assessment measures of achievement changes; 10 states showed changes that are more positive than the changes measured by NAEP, and 5 states showed changes that are less positive than those measured by NAEP.

- In grade 4 mathematics, 13 of 21 states showed no statistically significant difference between NAEP and state assessment measures of achievement changes; 5 states showed changes that

[4] Because differences between changes in achievement measured by NAEP and changes measured by the state assessment and the NAEP scale equivalents are based on the same data but are analyzed in different ways, statistically significant differences can be found in one and not the other because of the nonlinear relationship between scale scores and percentiles.

are more positive than the changes measured by NAEP, and 3 states showed changes that are less positive than those measured by NAEP.

- In grade 8 mathematics, 9 of 21 states showed no statistically significant difference between NAEP and state assessment measures of achievement changes, 7 states showed changes that are more positive than the changes measured by NAEP, and 5 states showed changes that are less positive than those measured by NAEP.

In considering the results described above, the reader should note that state assessments and NAEP are designed for different, though related purposes. State assessments and their associated proficiency standards are designed to provide pedagogical information about individual students to their parents and teachers, whereas NAEP is designed for summary assessment at an aggregate level. NAEP's achievement levels are used to interpret the meaning of the NAEP scales. NCES has determined (as provided by NAEP's authorizing legislation) that NAEP achievement levels should continue to be used on a trial basis and should be interpreted with caution.

In conclusion, these mapping analyses offer several important contributions. First, they allow each state to compare the stringency of its criteria for proficiency with that of other states. Second, mapping analyses inform states whether the rigor of their proficiency standards as represented by NAEP scale equivalents changed from 2005 to 2007. Significant differences in NAEP scale equivalents might reflect changes in state assessments and standards and/or other changes such as changes in policies or practices that occurred between the years. Finally, when key aspects of a state's assessment or standards remained the same, these mapping analyses allow NAEP to corroborate state-reported changes in student achievement and provide states with an indicator of the construct validity and generalizability of their test results.

CONTENTS

List of Tables

Table Page

Table Page

List of Figures

Figure Page

1 Introduction

State-level National Assessment of Educational Progress (NAEP) results are an important resource for policymakers and other stakeholders responsible for making sense of—and acting on—state assessment results. Since 2003, the National Center for Education Statistics (NCES) has been sponsoring research that focuses on comparing the proficiency standards of NAEP and states.[1] By mapping each state's standard for *proficient* performance onto the NAEP achievement scale, state policymakers can make comparisons of standards across states, in terms of the level of achievement required for *proficient* performance.

Recent studies that map state performance standards onto the NAEP scale have underlined the need for ongoing scrutiny of comparisons between NAEP and state assessment results.[2] In this report, we examine the consistency of the mapping results by using data from the state assessments and NAEP in 2005 and 2007. We investigate the impact and implications of the outcomes of the mapping procedure by using multiple years of data.

At a time when states are working to ensure that all their students reach proficient levels of achievement by 2014, as required by the No Child Left Behind Act of 2001 (NCLB), the analyses described in this report allow state policymakers to assess how high their state has set the bar for proficiency.

The comparison of achievement presented in this report is not intended to suggest deficiencies either in state assessments or in NAEP. The NAEP scales in reading and mathematics are being used as a common metric, not as a standard for evaluating state scales. Similarly, the NAEP achievement levels are provided simply as a national reference point for comparisons, not as a replacement for any given state's duly adopted state standards. Moreover, as provided by law, NCES, upon review of congressionally mandated evaluations of NAEP, has determined that NAEP achievement levels are to be used on a trial basis and should be interpreted with caution.[3] State-NAEP comparisons can help in the interpretation of state assessment results by providing a benchmark by which to assess changes in achievement that are measured by state assessments.

Using NAEP to compare state performance standards

The percentage of students identified as proficient on state assessments varies across states. Because each state's standard for proficient performance is set independently, the standards in different states can be quite different, even though they use the same terminology. A student who scores proficient in one state can move to another state and find that his or her performance is below the proficient range in the new state. NAEP, however, can provide the needed link to compare these assessment results across states. This comparison places all states' reading and mathematics standards on a common scale—the NAEP reading or mathematics scale—along

[1] Reports on this research are available at http://nces.ed.gov/nationsreportcard/studies/statemapping.asp.

[2] In early investigations, McLaughlin and Bandeira de Mello (2002, 2003, 2006) and subsequent reports (McLaughlin et al. 2008a, 2008b) mapped state primary performance standards onto the NAEP scale. Braun and Qian (2007) used similar methodology and data to conduct similar mappings. The recent mapping report from the National Center for Education Statistics is an outgrowth of these studies (U.S. Department of Education 2007).

[3] The status of NAEP achievement levels is available at http://nces.ed.gov/nationsreportcard/achlevdev.asp?.

with the NAEP achievement-level cut points. In this way, stakeholders can compare the relative stringency of state standards for proficiency in reading and mathematics.

A number of studies present arguments against the appropriateness of state assessment and NAEP comparisons. Criticisms of previous comparisons of state assessments and NAEP are not without merit and deserve thoughtful consideration. Prior criticisms of mapping studies have focused on three main topics: (1) state assessments and NAEP are developed for different purposes and have different goals and, as a result, should not be placed on a common scale; (2) state assessments may measure different constructs (e.g., Language Arts vs. Reading vs. Word Recognition) and should not be compared with one another; and (3) mapping studies implicitly use NAEP as the standard against which state assessments are ultimately determined to be deficient (Ho and Haertel 2007). Two National Research Council–sponsored studies have concluded that for a variety of reasons, mappings at the student level cannot be constructed validly (Feuer et al. 1999; Koretz, Bertenthal, and Green 1999).

Importantly, these studies do not address the appropriateness of mapping at the school level for the purpose of analyzing state-level results, which is the aim of the study described here. In a recent critique, Ho and Haertel (2007) posit that "substantial differences between state tests and NAEP will render the mapping illogical and subject to drift over time" (p. 1). If the standard that students must meet is not the same in 2007 as it was in 2005 (i.e., the standard "drifted"), then we cannot know whether achievement was better in one year or the other just because the percent achieving the standards was higher or lower in one year than in the other. Therefore, it is important to analyze NAEP and state assessment changes in achievement from 2005 to 2007, as was done for this report, which can determine whether there was drift. Drift indicates changes in either the state test or in NAEP (or in both) between consecutive administrations.

In an early mapping study, McLaughlin and Bandeira de Mello (2002, 2006) list a number of important caveats intended to prevent the misinterpretation of mapping results. They state emphatically that their report, among other things, (a) does not address questions about content, format, or administration of state assessments, as compared to NAEP, and (b) is not an evaluation of state assessments. As pointed out above, state assessments and NAEP are designed for different, although overlapping, purposes. For example, in many cases, state assessments are designed to provide pedagogical information about individual students to their parents and teachers, whereas NAEP is designed for summary assessment at an aggregate level. Findings of different standards, different trends, and different gaps should be presented without any implication that they be considered deficiencies either in state tests or in NAEP. However, it would be premature to conclude that two tests measuring grade 4 reading proficiency would assess no overlapping skills. Two tests that look quite different can measure the same variation because the various parts of reading (or mathematics) ability are highly correlated with one another. The high and consistent school-level correlations between state and NAEP assessment results suggests that state assessments and NAEP measure similar or related skills (McLaughlin *et al.* 2008, 2008b).

Despite the criticisms of NAEP and state assessment comparisons, there is a need for reliable information that compares state standards. What does it mean to say that a student is proficient in reading in grade 4 in Massachusetts? Would a fourth-grader who is proficient in reading in Wyoming be proficient in Oklahoma? However difficult it may be to answer these questions definitively, they are fair questions that deserve consideration. In this study we examine the consistency of state standards when mapped onto the NAEP scale from 2005 to 2007.

Data sources

The analyses in the report are based on NAEP and state assessment results of public schools that participated in NAEP, weighted to represent the states.[4] The analyses use data from these sources: (a) NAEP data files for the states participating in the 2005 and 2007 reading and mathematics assessments, (b) state assessment school-level files compiled in the National Longitudinal School-Level State Assessment Score Database (NLSLSASD), and (c) school-level achievement data for the 2006-07 school year from ED*Facts*.[5] This report also relies on a review of state assessment programs conducted to gain contextual information about the general characteristics of state assessment programs and to help identify changes in states' assessments between the 2004-05 and 2006-07 school years that could affect the interpretation of the mapping results.[6]

The analyses presented are based on the standard NAEP estimates, which do not represent the achievement of those students with disabilities and/or English language learners who are excluded from NAEP testing.

Organization of this report

The report presents mapping results using the 2005 and 2007 NAEP assessments in mathematics and reading for grades 4 and 8. The analyses conducted for this study addresses the following questions:

- How do states' 2007 standards for proficient performance compare with each other when mapped onto the NAEP scale?
- How do the cut points on the NAEP scale that are equivalent to the scores required to meet a state's standard in 2007 compare to those estimated for 2005?
- Using the 2005 NAEP scale equivalent standards to define a state's *proficient* level of performance on NAEP, do NAEP and that state assessment agree on the changes in the proportion of students meeting that state's standard for proficiency from 2005 to 2007?

Section 2 of this report provides a description of the estimation methods used in the mapping and in the comparisons of results between 2005 and 2007. Section 3 presents the results of the analyses that examined the mapping results for 2007 in reading and mathematics at grades 4 and 8. Addressing the second question, Section 4 focuses on the comparison between the 2005 and 2007 mappings in reading and mathematics at grades 4 and 8. Addressing the third question,

[4] The method for sampling private schools in NAEP precludes using private school results in state-related reports. All NAEP published statistics at the state level are therefore for public schools only. Also, because private schools are not required to participate in a state's annual academic assessments under NCLB, private school data are not generally included in state test score databases.

[5] ED*Facts* is a collaborative effort among the U.S. Department of Education, State Education Agencies, and industry partners to centralize state–reported data into one federally coordinated, K–12 education data repository, located in the U.S. Department of Education.

[6] State profiles based on the *2007 Survey of State Assessment Program Characteristics* are available at http://nces.ed.gov/nationsreportcard/studies/statemapping.asp.

Section 5 discusses the NAEP and state assessment changes in achievement from 2005 to 2007, including possible explanations for discrepancies in the gains measured by the state tests and NAEP so that attention can be turned to identifying the sources of those discrepancies. Tables in appendix A show the sample sizes and percentages of the 2007 NAEP samples used in the analyses. Tables in appendix B summarize selected changes in states' assessments between the two NAEP administrations of 2005 and 2007 that could affect the interpretation of the mapping results. Appendix C includes tables with results complementing those discussed in the body of the report.

2 Estimation Methods

State assessment scores are usually reported as percentages of students in a grade at a school whose test scores are sufficiently high to meet a predefined state standard. That standard has been shown to vary a great deal from state to state (McLaughlin and Bandeira de Mello 2003). As a result, comparisons of percentages of students meeting state standards in different states are as much, if not more, a function of the placement of the standards as they are of differences in the achievement of the students.

Of essence in any attempt to compare changes in achievement on two tests is an understanding that the increase in the percentage of students meeting the standard depends critically on the placement of the standard. Generally, standards placed near the median test score (or more specifically, the modal test score) show the most increase in percentages meeting the standard, whereas relatively high and low standards lead to smaller changes in percentages meeting the standard (McLaughlin and Bandeira de Mello 2003). However, there are exceptions to this generality. For example, if instruction focuses on a particular subgroup of students located at one end of the distribution, a standard set at that end may show larger changes than a standard set in the middle.

To account for variation in the placement of standards on a scoring scale, the first step in comparing NAEP and state assessment measures of change is to measure changes in NAEP performance the same way change is measured on state assessments, that is, using the percentage of students in the state meeting that state's standard.[7] The process is done by *mapping* each state's standard onto the NAEP scale; that is, finding the NAEP scale value for the NAEP sample in the state for which the estimated percentage of students with higher NAEP scale values matches the percentage of students reported by the state as achieving the state's standard in the same schools. Of course, because NAEP is based on a sample of students in each participating school, and because both assessments have measurement error, there is some mapping error in determining the NAEP equivalent of a state's standard. It is necessary to consider mapping error for valid comparisons between NAEP and state assessment measures of change in achievement.

This section summarizes the estimation methods used in the mapping procedure to place state performance standards onto the NAEP scales and in the comparison analysis between 2005 and 2007. We develop a framework for evaluating differences between achievement changes measured by NAEP and by state tests. Essentially, NAEP and state achievement changes in each subject and grade are rendered comparable by summarizing NAEP results in a state as the percentage meeting the state's standard, which requires, as a first step, mapping the state's standard onto the NAEP scale.

[7] Given that the only test results systematically available for all states are percentages of students in each school with scores higher than a cut point (i.e., *meeting the standard*), finding the NAEP equivalent of that cut point is an essential step in comparing achievement gains based on state test data to achievement gains on NAEP. If state test means and standard deviations were available for schools in the NAEP sample, mapping of the standards, while important in itself, would not be required for comparing state test and NAEP achievement gains.

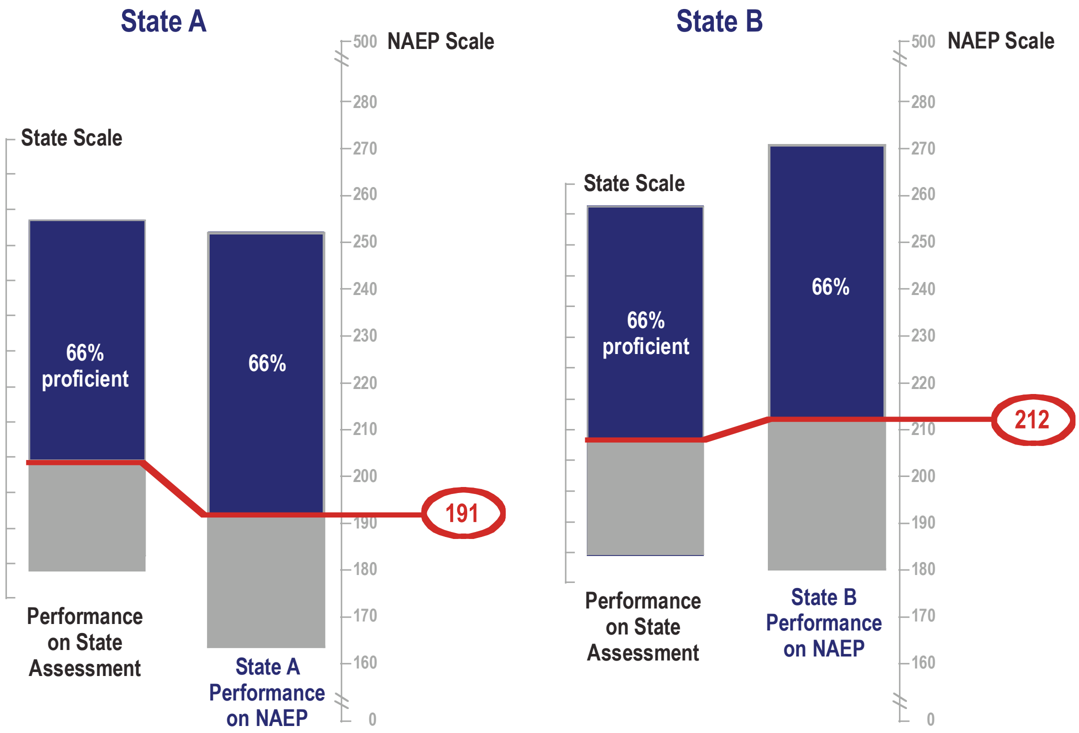
Estimation of the placement of state performance standards on the NAEP scale

The method of obtaining *equipercentile equivalents* involves the following steps:

1. Obtain for each school in the NAEP sample the proportion of students in that school who meet the state performance standard on the state's test.
2. Estimate the state proportion of students who meet the standard on the state test, by weighting the proportions (from step 1) for the NAEP schools, using NAEP school weights.
3. Estimate the weighted distribution of scores on the NAEP assessment for the state as a whole, based on the NAEP sample of schools and students within schools.
4. Find the point on the NAEP scale at which the estimated proportion of students in the state who score above that point (using the distribution obtained in step 3) equals the proportion of students in the state who meet the state's own performance standard (obtained in step 2).

Using figure 1 to illustrate, we see that 66 percent of the students in State A meet that state's standard (estimated from step 2); based on State A's NAEP sample, 66 percent of State A's students score above 191 on the NAEP scale (using the distribution obtained in step 3). Suppose that in State B, where students perform higher on NAEP than in State A, 66 percent of its students also meet its state standard. This translates into a higher NAEP scale equivalent (212 in the illustration), because 66 percent of State B's students score above 212 on the NAEP scale, based on State B's NAEP sample. State A's standard corresponds to, or maps onto, a lower level of NAEP achievement than State B's standard does, even though each state reports the same 66 percent meeting its own standard.

Figure 1. Mapping state proficiency standards onto the NAEP scale



The reported percentage meeting the state's standard in each NAEP school s, p_s (e.g., 66 percent as in figure 1), is used to compute a state percentage meeting the state's standards, p_S, using the NAEP school weights, w_s. For each school, w_s is the sum of the student weights, w_{is}, for the students selected for NAEP in that school.[8] For each of the five sets of NAEP plausible values, v = 1 through 5, we solve the following equation for c, the point on the NAEP scale corresponding to the percentage meeting the state's standard:[9]

$$p_S = \sum_{is,s\in S} w_{is} p_s \Big/ \sum_{is,s\in S} w_{is} \qquad [1]$$

$$= \sum_{is,s\in S} w_{is} \partial_{isv}(c) \Big/ \sum_{is,s\in S} w_{is} \qquad [2]$$

where the sum is over students in schools participating in NAEP, and $\partial_{isv}(c)$ is an indicator variable that is 1 if the vth plausible value for student i in school s, y_{isv}, is greater than or equal to c, and 0 otherwise. The five values of c obtained for the five sets of plausible values are averaged to produce the NAEP threshold corresponding to the state standard, that is, the reported mapping of the standard onto the NAEP scale.[10] Variation in results over the five sets of plausible values is a component of the standard error of the estimate, which is computed by following standard NAEP procedures.[11,12]

Relative error

When used to place state standards on the NAEP scale, equipercentile mapping will produce an answer even if NAEP and state assessment scores are completely unrelated to each other. Some additional data, beyond the percentage meeting the standard in the state and the distribution of NAEP plausible values—the only data used in the computation—are needed to test the validity of the mapping.

To evaluate the validity of the placement of a state standard on the NAEP scale, we measure how well the procedure reproduces the percentages reported by the state as meeting the standard in each NAEP-participating school. If the mapping is valid, the procedure should reproduce the individual school percentages fairly accurately. However, if the state assessment and NAEP are measuring different, uncorrelated characteristics of students, the school-level percentages meeting the state standard as measured by NAEP will bear no relationship to the school-level percentages meeting the state's standards as reported by the state.

[8] To ensure that NAEP and state assessments are equitably matched, NAEP schools that are missing state assessment scores (i.e., small schools, typically representing approximately 4 percent of the students in a state) are excluded from this process. Even if the small excluded schools perform differently from included schools, no substantial bias in the estimation process would be introduced, unless their higher or lower scoring was specific to NAEP or specific to the state assessment.

[9] Estimations of NAEP scale score distributions are based on an estimated distribution of possible scale scores (or *plausible values*), rather than point estimates of a single scale score. More details are available at http://nces.ed.gov/nationsreportcard/tdw/analysis/est_pv_individual.asp.

[10] Appendix A of McLaughlin et al. (2008a) describes in more detail the technical aspects of the placement of state achievement standards on the NAEP scale.

[11] NAEP computes standard error using a combination of *sampling error* based on Jackknife resampling and *measurement error* from the variance between plausible values.

[12] This mapping procedure is analogous to the one used in U.S. Department of Education (2007) and produces results that are qualitatively similar. The distinctions between the two procedures are discussed in Braun and Qian (2007).

The correlation coefficient showing the relationship between the percentages reported for schools by the state and those estimated from the NAEP scale equivalents provides a straightforward measure of the appropriateness of the mapping. However, it does not indicate the amount of error that is added to the placement of the standard by the fact that NAEP and the state assessment may not measure the same construct. We must determine how high the correlation must be to justify inferences that are based on the mapping. Also needed is a measure of that error, as a fraction of the total variation of percentages meeting the standard across schools.

The NAEP estimate of the percentage meeting the standard in a school is subject to both sampling and measurement error. However, even if the NAEP measure had no sampling or measurement error, and even if NAEP measured exactly the same construct as the state assessment, NAEP would not reproduce exactly the state assessment percentage for each school. The difference occurs because the state assessment scores are based on different administrations, at different times of year, with different motivational contexts and different rules for exclusion and accommodation. The state assessment scores are also subject to measurement error, although for school-level aggregates, the measurement error is smaller than it is for individual student estimates.

Although we recognize that discrepancies between the reported figure from each school and the estimate based on the NAEP mapping will occur, it is, nevertheless, important that the discrepancies be small relative to the variation in outcomes across schools. If the variance of the discrepancies is more than a fraction of the total variance across schools in percentage meeting a standard, the validity of the placement of the standard could be considered suspect, even though the nominal standard error of the state-level estimate may be small.

To evaluate the mapping, we therefore compare three variances:

1. total variance of reported percentages meeting the state's standard across the schools participating in NAEP in the state, $\sigma^2(p_s)$;
2. average squared deviation between the reported percentage, p_s, and the percentage based on the NAEP mapping for each school *s*, $\hat{p}_s$: $average_S\ (p_s - \hat{p}_s)^2$; and
3. average expected sampling and measurement error in the NAEP estimate for each school *s*, $average_S\ (\hat{p}_s - \mathrm{E}(\hat{p}_s))^2$.

We estimate the sizes of what the (squared) discrepancies would have been if NAEP were not subject to sampling and measurement error by subtracting quantity (*3*) from quantity (*2*), and we compare these adjusted (squared) discrepancies with the overall variation in percentages across schools $\sigma^2(p_s)$ (quantity (*1*)). If the adjusted (squared) discrepancies correspond to a large component the overall variance of the percentages, the NAEP data do not reproduce the school-level percentages with sufficient accuracy to justify inferences based on the placement of the standard on the NAEP scale. That is, we want the relative error $K < k$,

$$K = \left[\left(average_S\ (p_s - \hat{p}_s)^2 - average_S\ (\hat{p}_s - \mathrm{E}(\hat{p}_s))^2\right) \Big/ \sigma^2(p_s)\right] < k \qquad [3]$$

where $0 \le k \le 1$.

We want the discrepancy variance (*2*) to be less than a threshold *k* of the variance in the state test score school percentages (*1*), but we do not want to penalize the mapping for the measurement and sampling error in $\hat{p}_s$ (quantity *3*), which contributes to quantity (*2*). Therefore, we subtract (*3*) from (*2*) before dividing by (*1*). The resulting numerator of the relative error *K* is an estimate of the amount of discrepancy variance that cannot be accounted for by NAEP sampling and measurement error. Because both quantities (*2*) and (*3*) are sample estimates of variances, it is reasonable to expect that they will usually differ from the true variances of (*2*) and (*3*), and this can lead to (*2*) – (*3*) < 0 in some cases. In fact, if there were no linking error, we would expect (*2*) – (*3*) < 0 in half the cases, because (*2*) and (*3*) would be two estimates of the same variance.

Both the discrepancies and the estimation of NAEP random estimation error are more stable in schools with larger NAEP samples of students. Therefore, to increase the stability of the estimate of *K*, the average over schools was weighted according to the size of the NAEP sample of students in the school; a small number of NAEP schools with fewer than five NAEP participants are not included in the computations.

The NAEP random estimation error variance is the sum of two components, sampling error and measurement error. Because at the student level the variable of interest is a simple binomial variable (meets or does not meet the standard), to estimate the sampling variance we can use the binomial variance of the estimate of a percentage, $\hat{p}_s(100-\hat{p}_s)/n_s$, where n_s is the size of the NAEP sample in the school and $\hat{p}_s$ is the percentage of NAEP participants in the school with plausible values greater than the value estimated to be equivalent to the state standard. The binomial variance should be reduced by a finite population correction, $fpc = \sqrt{(N_s - n_s)/(N_s - 1)}$, because the NAEP sample is a sizeable fraction of the number of students in the particular grade, N_s, at most schools. If the number of students per grade is not known, the average finite population correction for schools with NAEP samples of the same size is used.

NAEP measurement error is estimated by the variance of the five estimates for each school's percentage meeting the standard, based on the five alternative sets of plausible values *v*, for the participating students, $\underset{v}{\sigma}^2(\hat{p}_{s,v})$. Because $\hat{p}_s$ is computed as the average of values based on five plausible value sets, the measurement error component is divided by 5. Thus, the quantity in (*3*) above is estimated by

$$\mathrm{E}\,(\hat{p}_s - \mathrm{E}(\hat{p}_s))^2 = (p_s q_s / n_s)(fpc)^2 + \underset{v}{\sigma}^2(\hat{p}_{s,v})/5\,. \qquad [4]$$

In this study, the criterion proposed is to consider relative errors greater than .5 as indicating that the mapping error is too large to support any useful inferences from the placement of the standard on the NAEP scale.

Setting the criterion for the validity of this application of the equipercentile mapping method at *K* = .5 is arbitrary but plausible. Clearly, it should not be taken as an absolute inference of validity—two assessments, one with a relative error of .6 and the other with .4, have similar validity. Setting a criterion serves to call attention to the cases in which we should consider a limitation on the validity of the mapping as an explanation for otherwise unexplainable results. Although estimates of standards with greater relative error because of differences in measures are not thereby invalidated, any inferences based on them require additional evidence. For example, a finding of differences in trend measurement between NAEP and a state assessment when the standard mapping has large relative error may be explainable in terms of unspecifiable differences between the assessments, ruling out further comparison. Nevertheless, because the relative error criterion is arbitrary, results for all states are included in the report and in the discussion of findings, irrespective of the relative error of the mapping of the standards.

Measurement error in comparing NAEP and state measures of change

Under No Child Left Behind, each state has developed measurements for determining whether its schools are making adequate yearly progress (AYP), which refers not to the progress of a child from, say, fourth grade to fifth grade but to the progress of a school in increasing the performance of its fourth-graders from one year to the next. The basic idea of comparing achievement changes from one year's students in a particular grade with achievement changes from another year's students in the same grade is that a set of skills is to be learned and that these skills might be more (or less) thoroughly learned by the students in one year than they were by the students in the other year. A test is written that samples the skill domain and is given to each of the two cohorts of students, and the scores are compared. Of course, the average scores will not be exactly the same in the two years if the test merely samples the skill domain and does so on a finite number of students. However, a simple statistical test can be executed to determine whether the difference is in the realm of random variation. If the sample of students were infinitely large and the test measured all the skills in the domain without error, the standard errors would be *zero*, meaning that any difference between the scores of the two cohorts would be statistically significant. Whether a difference is *important* is another question, but differences that are not statistically significant should not be considered further because they may well reflect just chance variation. [13]

Letting *D* be the discrepancy between changes from year 1 to year 2 in percentage meeting the state standard identified by the state test and the changes in the same period in the same percentages when measured by NAEP, we can test for whether *D* is statistically significantly different from zero by estimating the ratio of *D* to its standard error. However, to interpret the results of such a comparison, we also need to consider the explanations of statistically significant values of *D*. These discrepancies represent an additional source of error that contributes to the differences in achievement changes identified by NAEP and by the state assessment program. In general, such differences are hypothesized to be the result of some systematic difference between what the state assessment measures and what NAEP measures (in test content, student populations, or test administration). We call this a *true score error* to distinguish it from discrepancies arising from the finiteness of the samples and the imperfections of measurement.[14]

[13] The following discussion is excerpted from a report to NCES on the measurement error in comparing NAEP and state test gains (McLaughlin 2008).

[14] One source of error is due to the systematic differences in the domains of skills assessed by NAEP and the state assessment, and not to random measurement error or to sampling error. A second kind of error arises because both tests measure the domain with some error and because the mapping is based on a finite sample of students. The distribution of NAEP scores in the sample of NAEP students in the specified schools is likely to be slightly different from the hypothetical distribution of NAEP achievement of all students tested by the state in those schools, leading to small over- or underestimates of the NAEP scale equivalent of the state standard.

Measuring the standard error of D

Because the data available for mapping states' standards onto the NAEP scale are limited to school-level percentages of students achieving a state's standard in schools participating in NAEP, the critical statistic for comparing NAEP versus state-test score changes is

$$D = (\hat{p}_{2S} - \hat{p}_{2N|map=1}) - (\hat{p}_{1S} - \hat{p}_{1N|map=1}) \qquad [5]$$

where $\hat{p}_{YS}$ is the state percentage meeting the standard in year Y, estimated by the weighted average of the percentages in the NAEP schools, and $\hat{p}_{YN|map=1}$ is the percentage of the distribution of NAEP plausible values in the state in year Y, estimated by the (same) weighted average of the distributions in the NAEP schools, which are above the NAEP scale value that was found in year 1 to correspond to the state standard.

For example, if the state shows a gain from 50 percent to 60 percent meeting the standard and NAEP reports a gain from 50 percent to 55 percent meeting the state's standard, then
D = (60 – 55) – (50 – 50) = 5. The statistical question to be addressed is whether a value of 5 for *D* is larger than we would expect on the basis of measurement and sampling error.

The term in the second parenthesis of equation [5] is zero by definition, with no error, because the NAEP scale value onto which the state's standard is mapped (in year 1) is the value that forces an exact match of percentages (in year 1). That is not to say that $\hat{p}_{1S}$ and $\hat{p}_{1N|map=1}$ are error-free estimates of their respective population statistics, just that the second term in *D* is exactly zero. The errors in $\hat{p}_{1S}$ and $\hat{p}_{1N|map=1}$ contribute to the error in the other term $(\hat{p}_{2S} - \hat{p}_{2N|map=1})$ through mapping error.

Both NAEP estimates, $\hat{p}_{1N|map=1}$ and $\hat{p}_{2N|map=1}$, are based on percentages of the student score distribution meeting the same scale value, the one mapped from the year 1 data. To measure achievement changes in terms of percentages of students meeting a standard, it is necessary to use exactly the same standard for both years.[15] In fact, if achievement changes are measured purely in terms of percentages meeting a standard, finding *an achievement gain in the population* is equivalent to finding that *the test became easier for the population to meet the standard*. In other words, unless we are assured that the standard has not been lowered, we cannot infer that finding that the standard became easier for the population means that the population's achievement increased. We cannot exclude the possibility that the standard was lowered unless we have evidence to exclude it. An example of that evidence is finding that in both years, the standard is equivalent to the same NAEP score, if we assume that NAEP remained unchanged between the years. Thus, the question of whether NAEP and the state assessment agree on the size of achievement change is virtually equivalent to the question of whether the mapping of the state's standard onto the NAEP scale was stable over the two years.

Because the second term in the equation for *D* is zero, we can redefine *D* as

$$D = (\hat{p}_{2S} - \hat{p}_{2N|map=1}) \qquad [6]$$

[15] If we were to estimate $\hat{p}_{2N}$ from a mapping based on year 2 data, *D* would be identically zero, a meaningless result.

and focus on the estimation of the sources of error; that is, on the expected variation between *D* and the value it would take on if the estimates of the percentages meeting the standard were equal to their population values, $\hat{p}_{2S}$ and $\hat{p}_{2N|map=1}$.

Many factors contribute to random variation of *D* around its true value, which would be zero if NAEP and the state assessments show the same gains/losses.[16] However, in view of the complexity of any psychometric model for *D*, the most robust procedure for estimating the standard error of *D* is the standard NAEP procedure, combining NAEP measurement error, estimated by variation in values of *D* obtained for each of the five plausible value sets, with NAEP sampling error, estimated by the NAEP jackknife technique.

Measuring the standard error of the mapping

Estimating the standard error of the mapping is not a necessary step in determining the standard error of *D* because we can apply the NAEP jackknife technique directly to the estimate of *D*. However, an estimate of the standard error of the mapping is necessary to test the question of whether the NAEP scale equivalent of the standard is stable across the two years. If we denote the NAEP scale equivalent of the standard in year Y by $\hat{c}_Y$, then the standard error of the difference,

$$\hat{c} = \hat{c}_1 - \hat{c}_2, \qquad [7]$$

is just the square root of the sum of the squares of the standard errors of the two separate NAEP scale equivalents. That is,

$$SE(\hat{c}) = \sqrt{SE(\hat{c}_1)^2 + SE(\hat{c}_2)^2}\ . \qquad [8]$$

Each can be estimated by applying the NAEP jackknife technique to the mapping process.

Summary

The ultimate purpose for estimating the standard error of *D* is to decide whether differences between changes in achievement showed by NAEP and changes in achievement showed by the state are sufficiently large that they are not likely to be due to random factors. If the difference, *D*, is statistically significantly different from zero, students gained more on either the NAEP skill domain or on the state-specific skill domain than represented by those domains' contributions to variance in year 1. Focusing on state-specific content during instruction might be expected to lead to a positive value for *D*, whereas focusing entirely on NAEP content might be expected to lead to a negative value for *D*. Other explanations for a larger change on the state test exist. A statistically significant value of *D* may be due to a change in the content, administration, or scoring of either the state test or NAEP in the interval. For example, a change in the NAEP exclusion rates between years (for whatever reason) can lead to a significant *D*; a larger apparent state gain (i.e., a positive change) could be due to increased familiarity with and focus on the state test in the schools, with teaching students how to do particular kinds of items on the state test; and decreasing the focus on some aspects of NAEP content in the state curriculum between the two assessments could lead to a larger gain on the state test.

[16] These factors are discussed in McLaughlin (2008).

The key underlying assumption is that NAEP and the state assessment each remain essentially the same over the two years. If either test is substantively changed between the two years, then comparisons of changes identified on the two tests are not warranted. NAEP did not go through any substantive methodological changes between 2005 and 2007. However, in the years from 2005 to 2007, the focus of this report, many states changed their state assessments to ensure that they were complying with the regulations of the NCLB law, and finding values of *D* significantly different from zero in those cases is to be expected.[17]

It should be noted that the state assessment data available for this study include only a single number (percentages) reported for each school (for each subject and grade). *D* is based on a match of NAEP and a state's assessment at a single point in the state's achievement distribution.

Finally, there is the question of what is meant by stability of the mapping between two years and how it can be measured. In practical terms, the value of *D* is a measure of the *instability* of the mapping. If $D = 0$, the mappings in the two years yield identical results. If *D* is positive (the state showed a more positive change than the change measured by NAEP), that means that if we were to calculate the NAEP scale equivalent of the standard in year 2, the result would be a lower value on the NAEP scale than the equivalent obtained from the year 1 mapping. This does not necessarily mean that the state's standard got easier. If both NAEP and the state's assessment and scoring systems remained constant over the 2-year interval, it means that there were more gains on state-specific skills than on NAEP skills during the interval.

[17] Tables in appendix B summarize selected changes in states' assessments between the two NAEP administrations of 2005 and 2007.