EDUCATIONSECTOR
www.educationsector.org

Explainer

October 2007

Understanding NAEP: Inside the Nation's Education Report Card

By Margery Yeager

ACKNOWLEDGEMENTS

This publication was made possible by a grant from Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.

The author would like to thank Andy Rotherham, Thomas Toch, Kevin Carey, and Robin Smiles for their support and contributions to this project and several others who reviewed this report.

ABOUT THE AUTHOR

MARGERY YEAGER is the Chancellor's Critical Response Team Leader at the DC Public Schools and a former employee of the U.S. Department of Education. She produced this report during her rotation at Education Sector as a Presidential Management Fellow with the U.S. Department of Education.

ABOUT EDUCATION SECTOR

Education Sector is an independent education policy think tank devoted to developing innovative solutions to the nation's most pressing educational problems. We are nonprofit and nonpartisan, both a dependable source of sound thinking on policy and an honest broker of evidence in key education debates throughout the United States.

ABOUT THIS SERIES

Education Sector Explainers give lay readers insights into important aspects of education policymaking. They are not intended to be technical manuals.

1201 Connecticut Ave., N.W., Suite 850, Washington, DC 20036
202.552.2840 • www.educationsector.org

Testing has taken center stage in today's era of increased accountability in public education. But only one test promises to measure student achievement across the country, across demographic groups, and across decades: the National Assessment of Educational Progress (NAEP), often referred to as the "Nation's Report Card."

NAEP is a series of assessments in math, reading, and other subjects. It is given regularly to national samples of fourth, eighth, and 12th-grade students to determine both what they do know and what they should know. The National Center for Education Statistics (NCES), located in the U.S. Department of Education, administers NAEP, and the National Assessment Governing Board (NAGB), a bipartisan board composed of governors, state and local education officials, business leaders, teachers, principals, measurement experts, and parents, oversees and sets policy for the test. Both NCES and NAGB rely on testing contractors to develop, score, and report on the program. The Educational Testing Service (ETS), the testing giant responsible for the SAT, Advanced Placement exams, the Graduate Record Examination, the PRAXIS series used for teacher certification, and the English language test TOEFL, has been the primary NAEP contractor since 1983.[1] Other major contractors include Pearson, an international media and testing company, and Westat, a research corporation.

Since NAEP was created in 1969, it has become a trusted resource. Its scores are widely cited in the media to describe national achievement levels, trends, and gaps in student performance. The publication *Education Week* recently described the test as the "most influential research study and information source of the past decade."[2] NAEP data are also used by researchers and commentators as a proxy for evaluating the rigor of state standards and to assess educational progress under the federal No Child Left Behind Act (NCLB). In May, for instance, when NCES reported increases in NAEP history and civics results, observers used the information to tout the benefits of NCLB in improving student achievement across all subjects.

But what NAEP can and cannot tell us about student performance is often not well understood. The test design is technically complicated, leading to difficulty in interpreting and reporting its results. Scores, for instance, can not always be compared across grade levels or even across subjects. While a score of 240 on a fourth-grade reading test might indicate a student is proficient, the same score on an eighth-grade math assessment could mean the student is below proficiency. Such complexity leads to misinterpretations by the media and the public.

NAEP, moreover, is constantly changing. Like other standardized tests that influence policy, NAEP has been forced to expand its design and implementation to meet demands for more detailed information about the state of American education. What started out as a $1.9-million-a-year single measure of national student achievement is now an $88-million-a-year program, with multiple tests examining trends at the district, state, and national levels.[3] And policymakers and educators continue to call for NAEP's expansion—most recently proposing its use as a measure of curricula effectiveness, an anchor for other assessments, an accountability tool, and an international comparison benchmark.[4] Calls for further expansion persist even though the test is not designed to meet many of these objectives and cannot be expected to without a significant and costly overhaul.

And, despite its extensive use and valued reputation, NAEP is not without controversy. Testing officials have faced concerns about low participation rates among 12th-graders and the exclusion of some students with disabilities from testing. And a host of bodies have criticized the process testing officials use to create NAEP's achievement levels (basic, proficient, advanced) and how those levels are defined.

Still, NAEP remains an extremely important source of data, one of the only tools for reliably comparing student achievement across states and demographic groups and the only nationwide longitudinal assessment of student

achievement in the nation. Fully understanding both the mechanics of NAEP and its controversies is essential for policymakers, researchers, and practitioners seeking to know how students are currently performing in a range of academic subjects and how performance has changed over time.

This Education Sector Explainer discusses NAEP's origin and its expanding role, describes how the test is designed, how its scores are calculated and what those scores mean. It examines the controversies surrounding the reporting and use of NAEP data. And it examines the challenges facing the Nation's Report Card in a climate of relentless demands for more information on student achievement.

Making the NAEP

Today, NAEP has two primary goals: comparing student achievement across states and tracking changes in national educational achievement over time. To address this dual purpose, there are two corresponding NAEP tests. "Main" NAEP provides a biennial snapshot of student achievement nationally and in the states, while "long-term trend" NAEP measures changes in student achievement over time. *(See sidebar on main vs. long-term trend NAEP, Pg. 3.)* The main NAEP test is much more widely discussed because it tests a greater number of subjects and students. Unless otherwise noted, all references to NAEP in this report refer to the main NAEP testing series.

NAEP now measures student achievement for the nation, states, and 10 of the largest urban school districts. But this was not the case when the test was created in 1969. In 1965, federal lawmakers directed over $1.5 billion to K–12 education, via the landmark Elementary and Secondary Education Act (ESEA), a key part of President Lyndon Johnson's "Great Society" initiative and a precursor to today's NCLB. This enormous and unprecedented financial investment increased federal interest in measuring national performance and implementing accountability. But officials at the state and local levels resisted what they saw as federal intrusion into state policy. So NAEP was created with the agreement that information would *not* be reported for individual states or districts.

But state resistance to NAEP began to wane by the mid-1980s. After the 1983 report "A Nation at Risk" denounced the condition of American education, state leaders started to compare their assessment data to NAEP to show the impact of educational reform.[5] Federal officials also expressed interest in more information on student performance, and a study group was formed in 1986 to look into the matter. In their final report, the group, which was headed by Tennessee Gov. Lamar Alexander and former Spencer Foundation president H. Thomas James, noted that education is largely a state responsibility and emphasized the importance of state-level reporting.[6]

In response to the Alexander-James study and the increased state and federal interest in state achievement data, Congress, in the 1988 reauthorization of ESEA, authorized two trials of collecting and reporting state-level information. The trials, which began in 1990, were a success, with high voluntary state participation, strong support of most state officials, and two positive evaluations from the National Academy of Education (NAEd), a scholarly organization dedicated to advancing high-quality education research. And a once-reluctant Congress voted to continue the state tests with strong support from governors.[7]

As the interest in more fine-grained information continued, in 2002, the Council for Great City Schools (CGCS), a coalition of large urban school systems, led a push to expand NAEP reporting to provide results for major urban districts in addition to states. While districts were not prevented from participating before, the 2002 legislation appropriated funds to conduct the assessment and report results.[8] The Trial Urban District Assessment (TUDA) was first given in Atlanta, Chicago, Houston, Los Angeles, and New York in reading and writing in 2002. It was repeated and expanded in 2003 and 2005, to a total of 10 districts in reading and math. Selected districts also participated in the 2007 NAEP reading and math assessments earlier this year.

The Testing Population

Adding state and district data to NAEP made the test more useful, but it also created an enormous challenge: to provide reliable data on a public and private school population that includes more than 3 million students in each grade.

NAEP tests over 650,000 students in reading and math—a small fraction of the more than 50 million K–12 students in the United States. This tested group is a "sample," or

Table 1: Main NAEP and Long-Term Trend NAEP Comparison

	Main NAEP	Long-term Trend NAEP
The main NAEP testing series and the long-term trend NAEP use different questions, scoring standards, and student samples, but both tests include multiple choice and short and long-answer questions, and both provide results only for groups, not individual students. Both NAEP tests also provide information on variables that describe students, teachers, and schools, such as students' home life and peer groups, teachers' professional backgrounds, and school demographics, use of tracking, and provision of computers.		
Subjects tested	Reading, mathematics, science, writing, U.S. history, world history, geography, economics, civics, foreign language, and the arts	Reading and math
Grades/Ages tested	• Math, reading, writing, science, U.S. history, and geography grades 4, 8, 12 • Arts, grade 8 • Foreign language*, world history*, economics, grade 12	Reading and math, ages 9, 13, 17
Frequency	• Fourth- and eighth-grade math and reading, every two years • Twelfth-grade math and reading, writing, science, U.S. history, geography every four years • Other subjects every 8 years	Every four years
Content	Changes approximately once a decade to match changes in teaching practice and curriculum	Largely unchanged from initial administrations in 1971 and 1973
Sample	650,000 in 2005, samples in odd years are representative of states and selected urban districts	75,000 in 2004, provides only national data for major demographic groups
Scoring	• Reading, fourth- and eighth-grade math, history, and geography 0-500 scale • Science, writing, 12th-grade math, civics, 0–300 scale • Arts (music, theatre, and visual arts), each independently scored on a 0–300 scale	0–500 scale
Reporting	Uses achievement levels to report what students can and should be able to do • Basic—partial mastery of the knowledge and skills that are fundamental for proficient work • Proficient—solid academic performance, competency over challenging academic material • Advanced—superior performance	Uses performance levels to benchmark scores and document changes in performance over time • 150–350 in 50 point increments, each associated with a particular skill • E.g., math level 250 represents numerical operations and beginning problem solving

*These tests will be administered for the first time in 2012.

smaller set of students designed to be representative of the larger student population. NAEP is not designed to report the scores of individual students or schools. Rather, it reports the achievement of large groups of American students, such as those in a particular state, and subgroups by gender, race, and ethnicity.

In an average state, approximately 2,500 students from 100 schools are sampled from each subject and grade for NAEP.[9] States that receive federal aid for educationally disadvantaged students must participate in NAEP reading and math assessments for the fourth and eighth grades. State participation is voluntary in all other assessments. States must reach an 85 percent school participation rate in order to have their results reported because adequate participation is essential to producing score estimates that are representative of the student population.[10] The participation requirement also ensures that the scores are not biased by having only certain kinds of students taking the test.

These state samples are augmented in several ways to ensure they are representative of all students. First, if a state chooses not to participate in a NAEP testing, a

selected group of schools in the state will still be asked to participate for the national sample, but results will not be reported at the state level. Also, a separate and smaller national 12th-grade sample is added. State samples do not include these students because they tend to have lower participation rates. Finally, a national sample of private school students in grades 4, 8, 12 is added for purposes of comparison. Sampling weights, which ensure that data is adjusted to match the relative proportion of individuals in a population, are then used to make valid inferences from the sampled group to the larger population. (Long-term trend NAEP uses only a national sample and does not report state results, and therefore, uses a far smaller sample.)

Encouraging Participation

NAEP historically has been voluntary for students, schools, school districts, and states, but NCLB requires states to participate in the biennial fourth- and eighth-grade reading and math main NAEP assessments in order to receive their federal Title I funds, which benefit poor students. The outcome of NAEP testing has no impact on the amount of federal funds that states receive.[11] The federal government pays for all state NAEP administrations, so even prior to NCLB, more than 40 states participated each year.[12]

But student participation has been a greater challenge, particularly among 12th-graders, who often have lost interest in school by the second semester of their senior year when NAEP is administered. The "low-stakes" nature of the test has caused observers to question whether students, at all levels, are fully motivated to perform their best during NAEP testing. NAEP has none of the incentives or penalties associated with most other tests, such as end-of-course tests or high school exit exams that students must take and pass in order to advance or graduate. And on other tests, such as statewide assessments, the consequences are usually linked to teachers, administrators, and schools. Here, students receive their individual scores and often receive encouragement and support from parents, teachers, and administrators. In contrast, students are neither penalized nor rewarded for performance on NAEP, and students, parents, teachers, and administrators never learn individual or school results. Former NAGB chair Mark Musick has suggested that NAGB needs to "make a much more compelling case to students [to] do your best for your country."[13]

Historically, 12th-graders have participated at much lower rates than fourth- and eighth-graders, but the gaps between grades have been increasing. Twelfth-grade participation rates dipped to a low of 55 percent in 2002 in contrast to 79 percent for fourth-graders and 75 percent for eighth-graders. In 2005, 12th-grade participation rates were still stuck at 55 percent, but fourth-grade and eighth-grade rates were at 90 percent and 88 percent, respectively.[14] NAGB members have considered a variety of policy changes to boost participation, such as offering incentives, providing feedback on test performance, and moving the test earlier in the school year.[15]

But, in each case, practical or legal considerations or concerns about test validity have stood in the way of the changes being implemented. For instance, moving the test to the fall when seniors are more engaged would no longer be a fair measure of their knowledge and skills at the end of high school.

Studies suggest that the low-stakes nature of NAEP lowers student performance only slightly, if any. Researchers from the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) embedded a set of NAEP math questions in an eighth-grade state assessment to see if student performance would change on a higher-stakes test. The results showed a small effect for a few of the questions and researchers concluded that NAEP results may be affected slightly by decreased motivation, but that any effect is likely small.[16]

Another study published by CRESST found an effect among eighth-graders but not among 12th-graders. After asking student focus groups to suggest motivational rewards, researchers asked eighth- and 12th-grade students to answer NAEP items for a financial incentive ($1.00 for every correct answer), as part of a competition against other students, or as a personal challenge. Only the eighth-grade students who received a financial reward demonstrated any significant improvement in performance, and this effect was only on easy test items.[17] Twelfth-grade performance showed no difference, supporting the conclusion that low motivation does not significantly impair performance.

Inclusion and Exclusion

While it is tough to get some students to participate in NAEP testing, others are being excluded. In 2002, for

instance, roughly 40 percent of students with disabilities were excluded from testing, inciting outrage from special education advocates.[18] "We are deprived of essential information concerning the academic progress of children with disabilities and the quality of services they are receiving," said Jim McCormick, president of the Council for Exceptional Children.[19]

Overall, exclusion rates have been falling, dropping from 57 percent in 1992 to 35 percent in 2005.[20] But discrepancies in exclusion rates among states make it difficult for NAEP to provide consistent data across states. Variations in exclusion rates among states have also fueled concerns that higher exclusion rates lead to artificially higher state scores and an inaccurate picture of students' true achievement levels.

Today, the common practice in testing is to test virtually all students, including those with disabilities and language deficiencies, and to provide testing accommodations, such as additional time or Braille tests, to students who need them. And, since 1996, students with disabilities have been included in NAEP testing unless a school-level team determined a student could not participate even with accommodations or if a student's individualized education plan (IEP) entitled that student to testing accommodations not permitted on NAEP. NAEP allows common accommodations such as extra time, individual administration, or oral responses. But it prohibits those that interfere with skills that NAEP measures, such as using calculators on a portion of the math test or having the test administered via audiotape. Some accommodations cannot be provided for practical reasons. Testing, for instance, cannot be extended over multiple days since NAEP administrators are usually at school sites for only one day.[21]

English language learners (ELLs) are included in NAEP testing unless they have received fewer than three years of reading and math instruction in English and cannot demonstrate their performance, even with accommodations. ELL students are eligible for testing accommodations including extended time or oral presentation of test questions (except in reading), but may not take native language versions of the test.[22]

But simple differences in how states identify students in these populations can have a major impact on NAEP because testing samples are based on states' student categorization systems. If a state determines a student is an English language learner, for instance, NAEP will include his or her score in that student group. Another state might not consider the same student to be an ELL, and this score would then be excluded from the ELL category. Exclusion rates vary greatly across states, but more so in reading, for several reasons: inconsistent criteria for identifying students as ELLs or students with disabilities, differing interpretations of exclusion guidelines, and population differences and shifts, particularly with regard to the number of students identified as ELLs. Discrepancies also arise when one state has either stricter or more lenient policies for determining which students are excluded from participation altogether.

The NAEP Report Card

For more than 35 years, NAEP has highlighted education challenges, successes, and trends not evident through any other assessment. When the test was initially administered, results were only available nationally and by region, and it was difficult to determine exactly what the scores meant without any comparison points. It was evident, however, that scores were lower in the Southeastern United States, and overall scores were not as high as most policymakers and members of the public believed they should be.

During the 1980s, scores on the math and science long-term NAEP increased somewhat, while reading scores remained flat. Later, in the 1990s, main NAEP provided evidence of greater progress in states that adopted an early and aggressive approach to implementing standards-based education reforms. Both reading and math scores increased moderately during that decade, with more significant gains in math.

NAEP results suggest ongoing modest gains in math achievement after the signing of the federal No Child Left Behind Act in 2002. But NAEP has reported little to no increase in reading scores at the fourth- and eighth-grade levels. And recently, good news has been mixed with bad. While results show that Southeastern states are beginning to rise in state student-achievement rankings and African American students show evidence of narrowing racial achievement gaps in reading, 12th-grade reading and math scores have remained stubbornly stagnant over the past 15 years.

Today, NAEP also provides information on factors related to student achievement, either directly through NCES reports or by research conducted using NAEP data. The 2005 NAEP High School Transcript Study, for instance, found that students who take algebra in eighth grade or earlier are more likely to take advanced math classes. NAEP 1992 student questionnaires indicate that American students read very little outside of school. And 1988 research using NAEP data suggests that English fluency, not language spoken at home, is associated with the academic performance of English language learners (ELLs).

Since students with disabilities and ELL students have lower scores on average, state policies that consistently overidentify or unnecessarily exclude these students could affect overall results.[23] Researchers commissioned by NCES, for instance, have found that states that raise or lower their exclusion rates tend to have corresponding small increases or decreases in scores.[24]

For example, Kentucky excluded only 4 percent of its fourth-grade students in 1992 and 1994, but excluded 10 percent in 1998 and had a statistically significant gain in its reading results that year.[25] Based on this research, some critics argue that exclusion rates distort information about actual progress and give unfair advantages to states that exclude many students. And outside observers have charged that some states have been credited for artificial score increases due to increasing the number of students they exclude on tests.[26]

The percent differences among states in exclusion rates are significant. On the 2005 NAEP reading test, 14 percent of all Louisiana fourth-graders were excluded, compared to 2 percent in Wyoming and Alabama.[27] In California 33 percent of students are English language learners and 12 percent of these students were excluded; while in Texas, only 16 percent of students are ELLs, but 38 percent of these students were excluded.[28]

NAGB is largely unable to dictate how state, local, and school officials identify students with disabilities or ELL learners. The board did, however, vote in 2000 to alert readers to states with changes in exclusion rates of more than 3 percentage points from year to year. But the policy was rescinded soon after, when NCES reported it could not determine a specific value at which exclusion rates would have a significant impact on overall state scores.[29]

Testing and Scoring

In order for NAEP to be rigorous and comprehensive, (i.e. testing a large number of students in a variety of subjects as well as measuring an array of skills within a subject area), it must have a large set of test questions. For each test, several hundred questions are needed—far more than any one student has time to complete. Thus, students are tested in only one subject area and only take a small subset of the total test. Each test is broken up into 25-minute component blocks, and test booklets containing different two-block combinations are evenly distributed to the sample.

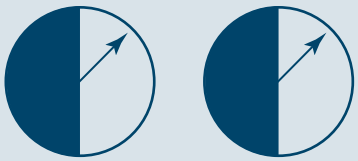
This process is designed to both reduce the burden on test-takers and schools and to provide results that are representative of all students. But the process increases the complexity in scoring and the reporting of results because each student is only exposed to a subset of the total test questions. To account for this, NAEP uses statistical techniques to compute a score that represents how each student would theoretically perform on the entire test.

NAEP uses a statistical methodology called "item response theory" (IRT) to predict how well students who did not see a particular question would have performed on it. Specifically, in this process a value called "theta" is derived for each question from three numbers: The first number is the question's difficulty—or probability of being answered correctly—for students at each of the three achievement levels (basic, proficient, and advanced). The second number is the chance that the question would be correctly answered by guessing, and this is dependent on whether it is a multiple choice, short answer, or long-answer question. The third number is the ability of the question to distinguish students of different ability levels. It is measured empirically by how well each student who answered that question performed on it and the other questions in the student's booklet.[30]

Once each question's theta has been determined, it is used to weigh each question appropriately in computing the score. For example, if a student answered many questions in a section wrong, but correctly guessed a difficult multiple choice question, item response theory would not weigh the multiple choice question highly in computing a scale score on the assumption that this response was a deviation from the student's overall pattern of responses.

As a result of this weighting process, and because students only take a small subset of the total test, numerical differences between scores do not correspond to any particular subset of questions answered correctly or incorrectly. Instead, score differences indicate a high probability that students could answer particular kinds of questions correctly. This complex scoring system helps NAEP make nuanced determinations of students' skills and abilities but also makes scores difficult for parents, teachers, and policymakers to understand.

Sample 12th-Grade Constructed-Response Math Question and Its Grading



The two fair spinners shown above are part of a carnival game. A player wins a prize only when both arrows land on black after each spinner has been spun once.

James thinks he has a 50–50 chance of winning. Do you agree?

A Yes **B No**

Justify your answer.

Solution:
Possible outcomes are BB, BW, WB, and WW.

Only BB will win. The actual chance of a win is 1 in 4, or 25%

Scoring Guide
In this question, a student has to determine the probability of a simple event. A student could have listed a sample space and used the information to describe and make a prediction about the expected outcome. Full credit was earned for a response that included the correct answer of 1 in 4, or 25%, with a complete justification (i.e., a list of the possible outcomes). Partial credit could have been earned if a student listed the sample space but the explanation was either incomplete or missing.

Score and Description

Correct
Correct response

Partial
Lists sample space correctly with less than a complete explanation

OR

draws a correct tree diagram with less than a complete explanation

OR

just states 1 in 4 chance

Incorrect
Incorrect response

Source: "NAEP, NQT v3.0" http://nces.ed.gov/nationsreportcard/itmrls/itemdisplay.asp.

Ensuring Accuracy

NAEP takes aggressive measures to ensure accurate and reliable scoring. Scoring is a huge challenge, because each test is given to thousands of students and contains both multiple choice and open answer questions. Multiple choice questions are electronically graded, but student-constructed responses, which are necessary for short answer questions, require a complex grading system to determine partial credit. (*See grading example sidebar, Pg. 7.*)

Scoring integrity is essential to providing valid and reliable test results, and NAEP uses rigorous scoring procedures to ensure accuracy and consistency in scoring. Items and scoring guides are developed by test contractors and thoroughly reviewed by NCES, NAGB, and state officials. All items are piloted before being used, and scoring guides are subsequently refined. NAEP scorers are required to have a bachelor's degree, sometimes in the specific subject area being graded. Scorers are trained on each item through scoring guides, examples of responses at each scoring level, and practice grading sessions. Scorers must also pass a qualifying test in which they are asked to score particularly challenging responses. To ensure quality, some responses are double scored to make sure that different raters will assign the same scores. Responses from previous years are also re-scored to ensure that grading is consistent across time.[31]

Interpreting the Scores

Most NAEP tests are graded on a 0-500 scale. This current scale was developed in 1984 and chosen to differentiate it from scales used for IQ tests, SAT scores, and grade equivalents.[32] But the scale used on NAEP is significantly harder to understand than those of other well-known tests, because the scores represent groups rather than individual students, and the scales for each subject are developed independently and cannot be compared.

This leads to misinterpretations by the media and public. Many people, for instance, assume that a score of 240 means the same level of performance for fourth- and eighth-graders. Yet, as the test is currently designed, numerical scores cannot always be compared across grade levels. Some NAEP subjects are scored on a "within-grade scale," which means the scores are derived

independently at each grade. Adding to the confusion, a few subjects are scored on a "cross-grade" scale in which scores are comparable across grade levels.[33] Changes in scales further complicate interpretation. For example, in 2005, the new 12th-grade math test used a 0–300 scale for the first time, which prevented any comparisons to data from previous years on the 0–500 scale. Such comparisons would not have been valid due to changes in the test's content.

Another basic problem is understanding the differences and changes in numerical scores. NAEP scores are not as simple to interpret as pure percentage scores or letter grades, e.g. 95 percent is an "A," 85 percent is a "B." A NAEP score of 220 is not 10 percent better than a score of 200, because there is no single formula to convert raw scores on test sections to scale scores for the test as a whole. Instead, the weight of each individual question in contributing to the scale score is determined by that year's student data. Additionally, changes in NAEP scores from one testing to the next may be only 1–2 points, but can be statistically significant due to the large sample size.

Reporting the Results

After each administration of main NAEP, officials publicly release the scale scores and achievement levels for groups of students. Scale scores are reported as averages—the average reading score for fourth-graders in 2005 was 219, compared to 218 in 2003. And achievement levels are reported as percentiles—in 2005, 64 percent of fourth-graders achieved at or above "basic" in reading, whereas 63 percent were at this level in 2003. Because understanding and interpreting the difference between 218 and 219 on an arbitrary scale is challenging even for informed observers, the more digestible achievement level percents usually receive wider play in the media. For example, a February 2007 press release on 12th-grade NAEP results noted declines in the percent of students scoring at or above proficient in reading and math, but made no mention at all of actual scale score levels.[34] At the same time, with a data set as large as NAEP, sometimes small changes in average scores have little practical significance.

In addition to achievement levels, each subject's results or "Report Card" includes an item map, which describes the particular skills associated with each achievement level as well as the progression of skills in between various scale scores.[35] *(See sample item map sidebar, Pg. 9.)* For example, on the 2005 NAEP eighth-grade reading test, a score of 247 is just above the "cut score," or passing score, for the basic level and indicates an ability to "locate specific information in a detailed document." A sample question demonstrating this skill level asks students to use a subway brochure to find factual information. A score of 318, which is just below the cut score for the "advanced" level, indicates an ability to "extend text information to generate a related question." Students might demonstrate this skill by posing a question to a character based on information in a reading passage.[36]

Misconceptions

Yet, even with the use of percentages and the item map, reporting the outcomes of NAEP is more difficult than other tests due to its complicated test design and scoring. The summary result reports provided by NCES attempt to provide clear information, but do not include many of the details of test design, sampling, and scoring, which are buried in technical reports or on the NAEP Web site. While intended to promote clarity, these omissions can often lead to incorrect assumptions about NAEP scores.

One common misconception is that a 10-point NAEP score difference represents one grade level. For example, on the 2005 reading test, NAEP achievement level cutoffs for basic, proficient, and advanced in eighth grade were approximately 40–50 points higher than the corresponding fourth-grade cutoffs. And some readers assumed that each 10-point score increment represented roughly one grade level in the scoring system. But this assumption is not accurate, as students in grades four, eight, and 12 take different tests. Also, each test is written in relation to standards for a particular grade and designed to measure a range of student achievement in that grade, not any other grades.

The media plays an important role in explaining NAEP reports to the public, and media reports on NAEP tend to focus on achievement levels more than scale scores, on the assumption that these categories are easier to grasp. However, this strategy tends to oversimplify reporting categories and does not recognize the span of scores within each category. The actual score difference between students in different achievement levels can be very large

or very small. Lumping students into categories masks the more discrete measure of performance indicated by scale scores.[37]

For example, after the release of the 2000 fourth-grade reading results, news stories noted an expanded gap between top performing and bottom performing students. Using percentile data, reporters offered incorrect descriptions of the drop in performance of students scoring at the 10th percentile. The *New York Times* erroneously described the decline as that of students in the below basic category (37 percent of all students).[38] The Associated Press and *Washington Times* had a different but, still, incorrect take on the information, suggesting that the decline represented the performance of all students in the bottom 10 percent.[39] Actually, the NAEP data showed that student scores had dropped for students scoring *at* the 10th percentile.

But even when reporters and observers clearly understand the meaning of NAEP scores, faulty interpretations are still possible. For example, the overall math performance of 17-year-olds on the "long-term trend" NAEP has not changed significantly since 1973, which could be interpreted as a lack of progress at this age level. But closer examination shows that white, black, and Hispanic students all showed improvement over this time period. The apparent contradiction is actually a result of the increasing percentage of Hispanic students taking the test, who, despite their improved performance, tend to score lower than white students.[40] This statistical phenomenon, known as Simpson's paradox, occurs when the trends of several groups seem to be reversed or negated when the groups are combined. This occurs because of a hidden variable—in this case relative size of each demographic group—which becomes influential when the data are combined.

Despite the difficulties of conveying accurate and clear information to the public, NAGB has been reluctant to use a simpler scoring system. And even simpler systems, such as percentage of answers correct, can be subject to interpretation problems. Percentages appear simpler, but they are potentially even more misleading. On the fourth-grade reading test, for instance, answering approximately 38 percent correct equates with a basic level, and roughly 62 percent is considered proficient.[41] This is a far cry from tests that are generally given in schools on which a 62 percent would be a failing grade.

Scores and Skills: Sample "Item Map" (2005 NAEP Reading, Eighth Grade)

500	
360	
356	Provide and explain evaluation of a document
350	
340	
336	Use examples to compare poetic language to everyday speech
332	Negotiate dense text to retrieve relevant explanatory facts
330	
327	Explain action in narrative poem with textual support
325	Provide specific explication of poetic lines
323	Explain the meaning of an image in a poem
323	**Advanced**
320	
318	Extend text information to generate related question
310	
301	Describe difficulty of a task in a different context
300	Provide support for judgment
300	
299	Recognize author's device to convey information
297	Recognize meaning of poetic comparison
295	Use metaphor to interpret character
290	
284	Apply text information to hypothetical situation and explain
284	Recognize what story action reveals about character
281	**Proficient**
280	
279	Relate text information to hypothetical situation
278	Infer character's action from plot outcome
275	Use task directions and prior knowledge to make a comparison
270	
267	Provide supporting details to explain author's statement
262	Use context to identify meaning of vocabulary
261	Identify causal relation between historical events
260	Identify appropriate text recommendation for a specific situation
260	
254	Explain reason for major event
253	Make inference based on supporting details to identify feeling
250	
248	Recognize information included by author to persuade
248	Provide specific text information to support a generalization
247	Locate specific information in detailed document
243	**Basic**
240	
237	Recognize significance of article's central idea
234	Provide partial or general explication of poetic lines
232	Identify characterization of speaker in poem
230	
228	Recognize an explicitly stated supporting detail
220	
210	Identify appropriate description of character's feelings
210	Identify main topic of informational passage

Source: "The Nation's Report Card: Reading 2005," US. Department of Education, NCES, 2006-451 (2006), available online at http://www.nces.ed.gov/nationsreportcard/pdf/main2005/2006451.pdf.

The Achievement Level Controversy

NAEP's achievement levels (basic, proficient, advanced) are more than just useful reporting mechanisms. When a newly appointed NAGB designed the levels in 1990 as a guide for what students should know and be able to do, with "basic" representing partial mastery of a subject, "proficient" representing solid performance, and "advanced" representing superior performance, it marked a key transition for NAEP.[42] The test shifted from strictly *reporting* performance, to *judging* performance against a standard, thus, expanding its focus in measurement to including evaluative and interpretive functions.[43] The goal of this change was to communicate clearly students' academic performance against an external standard. Says research professor and former NAGB member Diane Ravitch: "No single aspect of NAEP has been more valuable to the public … nor more controversial."[44]

The transition led to immediate criticism as well as spirited defense of the standards. Critics argue that the levels give a false sense of accuracy, when, in fact, they, like all standard-setting processes, are inherently subjective. And each time officials use NAEP results to make claims about the state of American education, a chorus of critics denounces the achievement levels, often suggesting they are artificially high and politically motivated. For instance, critics recently attacked a report on the effectiveness of education in each state issued by the Center for American Progress and the U.S. Chamber of Commerce, arguing that the report's conclusions relied on "misleading information" from NAEP achievement levels.[45]

On the other hand, those involved in the standards process argue that it is a carefully implemented approach that suffers from consumers wanting to attribute more to the results than the results are able to provide. Mark Reckase, a consultant to NAGB, describes the standard-setting process as "the most thoroughly planned, carefully executed, exhaustively evaluated, completely documented, and most visible of any standard-setting process" he has encountered.[46]

At the center of the controversy is the process officials use to create such performance targets. NAGB used a modified "Angoff method" to determine the "cut" or passing scores for each level. Here, panels of teachers, business leaders, state and local education officials, and testing experts evaluate test questions to determine the probability that a student just reaching each achievement level could answer the item correctly. Their collective responses are then averaged to determine a cutoff point for each achievement level.[47] Later, the panels evaluate the test as a whole, instead of individual test questions, and adjust the achievement level thresholds accordingly.

Yet, this method, like all standard-setting processes that must rely at some point on human judgment, is inherently subjective, leading to broad disagreement among researchers. As education professor Edward Haertel aptly notes, there is no "right answer waiting to be discovered."[48]

But many researchers have questioned NAEP's standard-setting methodologies. A series of evaluations throughout the 1990s excoriated the achievement levels and the process used to set them. The federal Government Accountability Office (GAO) commented in 1993 that "NAGB's approach [to setting achievement levels] is unsuited for NAEP," and characterized the resulting levels as "misleading."[49] And when the federal National Academy of Science (NAS) analyzed the development of achievement levels for the 1996 science test, they, like earlier researchers, concluded that the process was "fundamentally flawed" because of the difficult and confusing task given to judges, inconsistencies in their judgments of items, lack of evidence for cut scores, and the unreasonable results that came out of the process.[50]

The latest research to fan the flames of the achievement level controversy is a study that compares the NAEP achievement levels to student scores on the Trends in International Mathematics and Science Study (TIMSS), an international math and science assessment administered to students in over 40 countries. The study found that in 2003, while only 26 percent of American students scored proficient or higher on the 2000 NAEP math assessment, even the top-scoring TIMSS country, Singapore, had only 73 percent of its students achieve "proficient," according to the NAEP levels. Similarly, in science, where only 31 percent of American students scored proficient on the 2000 NAEP science assessment, just 55 percent of students in Singapore achieved this level. Critics assert that if the top country in the world can't achieve 100 percent proficiency in math or science, the standard is unreasonably high.[51]

Critics also question the political motivations behind the inclusion and ongoing use of NAEP's achievement levels.

Some have suggested that the levels were designed to change NAEP's role from a simple thermometer to a political tool to catalyze reform.

Defenders of the achievement levels note that there is a tension between what the data supports and what some stakeholders want NAEP to provide. The achievement levels were previously defined around providing information on what students at various achievement levels could do, such as an advanced level demonstrating "readiness for rigorous college courses." But NAGB revised the definitions of the levels in response to concerns that this could not be tied to the data. For example, without individual level scores, it was difficult to prove that students scoring at the advanced level were, in fact, later successful in college classes. Instead, the levels now provide greater technical accuracy in their descriptions but may have definitions that provide less clarity to its audience.

In 1994, when Congress considered ESEA reauthorization, the controversy surrounding the achievement levels reached such a point that the House Education and Labor committee voted to abolish the NAGB.[52] Yet, thanks to allies among governors and federal officials, the full House bill reinstated NAGB, though in weaker form and without authority to set achievement levels. The Senate, however, ultimately rose to NAGB's defense, restoring its full authority in the final law. But lawmakers did include in the 1994 ESEA reauthorization language describing the achievement levels as "developmental." And later, the 2001 reauthorization of ESEA as NCLB contained a similar provision specifying that the controversial achievement levels should be used on a "trial" basis until determined through evaluation to be "reasonable, valid, and informative to the public."

Yet, despite the criticism of NAEP's achievement levels, they continue to be widely used in reporting NAEP results. And critics also continue to assert that political motivations explain the lack of revisions to the method or actual achievement levels used. Gerald Bracey, an education commentator and ardent NCLB and NAEP critic, suggests that "much political hay can be made by alleging that American students are performing poorly."[53] NAGB has responded to this kind of criticism by affirming the importance and utility of achievement levels and noting in NAEP materials that "a proven alternative to the current process has not yet been identified."[54]

When asked recently about the ongoing controversy surrounding NAEP's achievement levels and the process used to create them, Chester Finn, a strong advocate of school reform who chaired the NAGB during much of the standard-setting process, recalled that the board agonized for many months over how many levels to set, what to call them, and how to set them. "There's no perfect way to set them; that was clear," he said. But he also addressed critics who continue to fault the current process: "What the critics of the standard-setting methodology have failed to appreciate is that at day's end, setting standards for educational performance is not a scientific or technical act; it's an act of judgment. And, at day's end, NAGB made those judgments—and still does. I'm proud of the judgments we made and the information that they yielded about American educational performance."

How the achievement levels are determined and labeled continues to have a strong pull on public policy by influencing public perceptions of how well public school students are performing. If the message is that the education system is failing and few students are reaching "proficient," the policy consequences are likely to be drastically different from those if the message is that most students are "proficient" and have the skills they need to succeed.

Using the Data

NAEP's expansion to include state-level data in 1990 and district-level data in 2002 created a higher profile for the test, but also has led to a greater potential for misuse.

Despite improvements to the test, NAEP has some serious limitations that impede its ability to provide comprehensive information to policymakers. State-level NAEP, for instance, provides snapshot data of achievement at a particular point in time for grades four and eight. It does not track a single set of students over time, which makes it difficult to use in tracking the outcomes of education reforms and policies, or to measure individual school performance.

Yet policymakers, eager to cite NAEP as evidence of the success of education reform, are quick to provide specious interpretations that cast favorable light on their state or district. A 1996 study by researcher and author Richard M. Jaeger found that achievement differences

across states were erroneously equated as differences in school quality, statistically insignificant differences were interpreted as meaningful, and public officials were likely to offer unsubstantiated causal explanations for changes in student performance on NAEP.[55]

Unfortunately, NAEP is widely used to support or oppose educational reforms, a role for which it is ill-suited. For example, in 2004, the American Federation of Teachers issued a report using NAEP data that was covered on the front page of the *New York Times*, which asserted that the performance of children in charter schools was inferior to that of students in public schools, implying that charter schools were a failing reform.[56] The U.S. Department of Education issued a report later that year noting that there were few differences in performance between students in public and charter schools, but that the math achievement of students in charter schools lagged their public school peers. But NAEP provides only a picture of current student performance and does not measure how much schools teach students in a year. Therefore, while this initial study provided a snapshot of how charter school students were performing at the time, NAEP was not able to provide much information on how much charter schools improved student performance.

Many outside organizations and observers use NAEP to validate the rigor of state tests. Advocacy organizations such as Achieve, Inc. and the Thomas B. Fordham Foundation, headed by Chester Finn, publish reports highlighting "proficiency gaps," or differences between proficiency rates on state tests and proficiency rates on NAEP, to argue that state standards are set too low. In general, far more students are deemed proficient on state tests than on NAEP, with differences as large as 60 percentage points.[57] A recent NCES report found that states vary widely in the gaps between their state proficiency rates and NAEP proficiency rates. For example, in fourth-grade reading, Massachusetts has the smallest gap with 48 percent of students meeting state proficiency standards and 44 percent meeting NAEP standards. At the other end of the spectrum, in Mississippi, 88 percent of students meet the state proficiency standard while only 18 percent meet the NAEP proficiency mark. Overall, the study found that most states set proficiency targets that fall in the basic range on the NAEP scale.[58]

Recent research from the Center on Education Policy indicates that NAEP frequently does not confirm the results of state tests, and presents a largely bleaker picture. In an analysis of state test data since 2002, researchers found that states that showed gains in student achievement on state tests generally did not experience similar gains on NAEP. Overall, gains on state achievement tests far outpaced the small score gains on NAEP.[59] The "proficiency gaps" revealed by NAEP have become fodder for those who advocate national standards and testing, and who argue that NAEP results show that states can't be trusted to hold the line on accountability.

In January 2007, two Democratic senators introduced bills that embrace voluntary national standards. In each bill, NAEP or its oversight board, NAGB, would have a major role. Senator Christopher Dodd (D-Conn.) introduced a bill that would require NAGB to develop voluntary national standards in math and science and provide grants to states to adopt those standards. Dodd's bill would also expand NAEP to test reading, math, and science in grades four, eight, and 12 every two years. A competing bill introduced by Senator Edward Kennedy (D-Mass.) would provide funding for groups of states to establish common standards and tests benchmarked against NAEP. Kennedy's bill would also require NAEP to ensure its standards are internationally competitive and expand 12th-grade testing to include a measure of college and workforce readiness.[60]

In July 2007, Senators Joe Lieberman (I-Conn.), Mary Landrieu (D-La.), and Norm Coleman (R-Minn.) introduced a bill that would require NAGB to work in conjunction with local, state, and national leaders to develop voluntary national standards and assessments for reading, math, and science. As an incentive, these standards and assessments would be provided for free to states willing to implement them, thus freeing up large amounts of state resources for other educational priorities. This legislation was endorsed by the Aspen Commission on NCLB, which had put forth a similar proposal in its February 2007 report.[61]

But there are important differences between state tests and NAEP that can generate differences in test results. The criteria for exclusion of students with disabilities and ELLs differ from state tests to NAEP. And in order to meet NCLB accountability provisions, states must test nearly all of their students on state assessments. In contrast, NAEP excludes roughly 5 percent of all sampled students, and

approximately 40 percent of all sampled students with disabilities. Since the students who are excluded tend to have lower average scores, some observers suggest that this could artificially inflate NAEP scores relative to state assessments. Yet, the majority of state assessments report higher percentages of students performing at the proficient level than NAEP does.

And there are other features of state tests that may account for part of the performance gaps with NAEP. State tests are high stakes tests with consequences for schools and sometimes for students too, which may have some limited effects on student achievement. State tests also are aligned with state standards and curriculum and may more accurately reflect student achievement if skills are taught in a different order than in NAEP frameworks. Consider subjects such as math and science, where subtopics may be designed to build upon one another. Here, a state's own standards may not align with NAEP frameworks and thus provide an inaccurate picture of student achievement.

A National Tool

NAEP remains one of the few national tools we have to gauge the overall effectiveness of American schools. NCLB focuses on state assessments, which vary widely in their content, alignment, rigor, and cut-score levels. It is extremely difficult to use them to make comparisons across states or to provide national averages. Other national tests also have limitations; tests like the SAT or ACT college admission tests only assess a specific segment of the student population, and only at the high school level. These tests are not nationally representative, and do not provide achievement levels, or a judgment of what knowledge and skills students *should* have at various grade levels.

And while it is virtually impossible for a single test to provide all the information needed to craft education policy, NAEP, with its recent commitment to release math and reading data within six months of test administration, is an increasingly timely source of student achievement information for the education community at the federal, state, and district levels. NAEP also disseminates information collected on student, teacher, and school background, which can be used to inform parents, the public, and policymakers about the impact of educational reform efforts in the nation's schools.

NAEP's limitations, however, should not be overlooked in the quest to find better information about the performance of the nation's schools. Nor should the test's sophisticated design and scoring system be underestimated. Only when NAEP results are accurately interpreted and reported can they best serve educators, parents, researchers, policymakers, and, most importantly, the nation's students.

Endnotes

[1] "The History of NAEP Partners," National Center for Education Statistics, U.S. Department of Education, http://nces.ed.gov/nationsreportcard/contracts/history.asp.

[2] Christopher B. Swanson and Janelle Barlage, Influence: A Study of the Factors Shaping Education Policy (Washington, DC: Editorial Projects in Education Research Center, December 2006).

[3] Maris A. Vinovskis, Overseeing the Nation's Report Card: The Creation and Evolution of the National Assessment Governing Board (NAGB) (Washington, DC: U.S. Department of Education, 1998).

[4] James W. Pellegrino, Lee R. Jones, and Karen J. Mitchell, Eds. Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress (Washington, DC: National Academy of Sciences, 1999).

[5] The National Commission on Excellence in Education, A Nation at Risk: The Imperative for Educational Reform (Washington, DC: U.S. Department of Education, April 1983). A Nation At Risk was a landmark report that decried the poor state of American education and served as a major catalyst for education reform.

[6] Lamar Alexander and H. Thomas James, eds., The Nation's Report Card: Improving the Assessment of Student Achievement (Washington, DC: National Academy of Education, 1987); Maris A. Vinovskis, "Overseeing the Nation's Report Card."

[7] Lynn Olson, "Governors Urge NAEP Expansion to Compare States," *Education Week*, February 13, 1991.

[8] Lynn Olson, "Budget Makes NAEP Testing Possible for 5 Urban Districts," *Education Week*, January 30, 2002.

[9] NAEP, "How the Samples Are Selected," National Center for Education Statistics, U.S. Department of Education, available online at http://nces.ed.gov/nationsreportcard/about/nathow.asp.

[10] NAEP, "State Assessment Sample Design FAQ," National Center for Education Statistics, U.S. Department of Education, available online at http://nces.ed.gov/nationsreportcard/about/samplesfaq.asp.

[11] "Important Aspects of No Child Left Behind Relevant to NAEP," National Center for Education Statistics, U.S. Department of Education, available online at http://nces.ed.gov/nationsreportcard/nclb.asp.

[12] "History of State Participation, 1990–1998: Public Schools," National Center for Education Statistics, U.S. Department of Education, available online at http://nces.ed.gov/nationsreportcard/about/statehistorypublic.asp.

[13] Sean Cavanagh, "Board Studies Release of Individual NAEP Results," *Education Week*, March 16, 2005.

[14] James R. Chromy, Participation Standards for 12th-Grade NAEP (Research Triangle Park, NC: RTI International, November 2005). Rates refer to combined school and student national response rates on NAEP reading before substitution.

[15] Sean Cavanagh, "Governing Board Looks to Marketing to Sell NAEP to Seniors," Education Week, November 24, 2004; Sean Cavanaugh, "Board Studies Release of Individual NAEP Results."

[16] Vonda L. Kiplinger and Robert L. Linn, Raising the Stakes of Test Administration: The Impact on Student Performance on NAEP (Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), March 1993).

[17] Harold F. O'Neil, Jr., Brenda Sugrue, Jamal Abedi, Eva L, Baker, and Shari Golan, Final Report of Experimental Studies on Motivation and NAEP Test Performance (Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), June 1997).

[18] Marnie S. Shaul, "NAEP Exclusion Rates for Students With Disabilities," GAO-06-194R, Government Accountability Office, October 28, 2005.

[19] "The High Number of Students With Disabilities Excluded from Federal Assessment Unacceptable, Says CEC," Council for Exceptional Children Press Release, November 10, 2005.

[20] Debra Viadero, "GAO Revises Estimates of Students Excluded from NAEP," E*ducation Week*, November 9, 2005.

[21] "Parents Frequently Asked Questions About NAEP," National Center for Education Statistics, U.S. Department of Education, available online at http://nces.ed.gov/nationsreportcard/parents/faq.asp.

[22] "Inclusion of Special-Needs Students: Criteria," National Center for Education Statistics, U.S. Department of Education, available online at http://nces.ed.gov/nationsreportcard/about/criteria.asp.

[23] Judith Anderson Koenig and Lyle F. Bachman, eds., Keeping Score for All: The Effects of Inclusion and Accommodation Policies on Large-Scale Educational Assessment, Executive Summary (Washington, DC: National Academy of Sciences, 2004).

[24] NAEP, "2005 Investigation of the Potential Effects of Exclusion Rates of Assessment Results," National Center for Education Statistics, U.S. Department of Education, available online at http://nces.ed.gov/nationsreportcard/about/2005_effect_exclusion.asp.

[25] David J. Hoff, "Board Won't Revise State NAEP Scores," Education Week, May 19, 1999.

[26] Kathleen Kennedy Manzo, "NAEP Exclusion Rates Increase for Disabled and LEP Children," *Education Week*, July 9, 2003.

[27] Debra Viadero, "States Vary on Students Excluded From NAEP Tests," Education Week, November 2, 2005.; NAEP Reading Mathematics 2005, "Reading State Exclusion Rate Tables," National Center for Education Statistics, U.S. Department of Education.

[28] "Nation's 'Report Card' Called a Faulty Instrument," California School News, California School Board Association, November 2005.

[29] Lynn Olson, "NAEP Board Worries States Excluding Too Many From Tests," *Education Week*, March 19, 2003.

[30] Robert J. Mislevy, Eugene G. Johnson, and Eiji Muraki, "Scaling Procedures in NAEP," Journal of Educational Statistics 17, No. 2 (Summer 1992): 131–154.

[31] "NAEP Item Scoring Process," National Center for Education Statistics, U.S. Department of Education, available online at http://nces.ed.gov/nationsreportcard/contracts/item_score.asp.

[32] Richard M. Jaeger, "Reporting the Results of the National Assessment of Educational Progress," (paper commissioned by the NAEP Validity Studies (NVS) Panel, September 1998).

[33] "Interpreting the 2005 Mathematics Results," and "Interpreting the 2005 Reading Results," NAEP, Institute of Education Sciences, U.S. Department of Education, available online at http://nces.ed.gov/nationsreportcard/mathematics/interpret-results.asp; NAGB, "Reading Framework for the 2009 Reading National Assessment of Educational Progress," available online at http://www.nagb.org/frameworks/reading_fw_08_05_prepub_edition.doc.

[34] "High School Students Show No Progress in Reading, According to the Nation's Report Card," National Assessment Governing Board Press Release, February 22, 2007.

[35] Links to each subject's most recent Report Card can be found at http://nces.ed.gov/nationsreportcard/.

[36] Marianne Perie, Wendy S. Grigg, and Patricia L. Donahue, National Center for Education Statistics, Institute for Educational Sciences, U.S. Department of Education 2005. "Nation's Report Card: Reading 2005," available online at http://nces.ed.gov/nationsreportcard/pdf/main2005/2006451.pdf.

[37] Daniel Koretz and Edward Diebert, Interpretations of National Assessment of Educational Progress (NAEP) Anchor Points and Achievement Levels by the Print Media in 1991 (Santa Monica, CA: Rand Corporation, 1993).

[38] Kate Zernike, "Gap Between Best and Worst Widens on U.S. Reading Test," *New York Times*, April 7, 2001.

[39] Andrea Billups, "Reading Test Find Fourth-Graders Stalled," *The Washington Times*, April 7, 2001; Greg Toppo, "Study: Fourth Graders' Reading Lags," Associated Press, April 6, 2001.

[40] NAEP Newsroom, "Long-Term Trend FAQ," National Center for Education Statistics, U.S. Department of Education.

[41] Nancy L. Allen, John R. Donoghue, and Terry L. Schoeps, "The 1998 NAEP Technical Report" National Center for Education Statistics, U.S. Department of Education, June 2001.

[42] The mandate to create achievement levels came in the 1988 ESEA reauthorization, which called for the development of "appropriate achievement goals" for each NAEP subject. The 1988 reauthorization also created the NAGB to oversee NAEP.

[43] Nancy L. Allen, John R. Donoghue, and Terry L. Shoeps, "NAEP 1998 Technical Report."

[44] Diane Ravitch, "Introduction," in Brookings Papers on Education Policy, (Washington, DC: Brookings Institution Press, 2001), 6.

[45] Gerald W. Bracey, "A Test Everyone Will Fail," *The Washington Post*, May 3, 2007, A25.

[46] Mark Reckase, "The Controversy Over the National Assessment Governing Board Standards," in Brookings Papers on Education Policy, ed. Diane Ravitch (Washington, DC: Brookings Institution Press, 2001), 231.

[47] NCLB required that these achievement levels be used on a trial basis until the Commissioner of Education Statistics determines that the levels are "reasonable, valid, and informative to the public." See NAEP Standards: http://nces.ed.gov/nationsreportcard/achlevdev.asp?id=ma%22; For more about how states set cut scores and for a more detailed discussion of the Angoff Method, see Andrew J. Rotherham, Making the Cut: How States Set Passing Scores on Standardized Tests (Washington, D.C.: Education Sector, July 2006).

[48] Edward Haertel, "Comment," in Brookings Papers on Education Policy, p. 256.

[49] General Accounting Office, "Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations," Report 93–12, June 1993.

[50] James W. Pellegrino, Lee R. Jones, and Karen J. Mitchell, Eds. Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress.

[51] Gary W. Phillips, Expressing International Educational Achievement in Terms of U.S. Performance Standards: Linking NAEP Achievement Levels to TIMSS, (Washington, DC: American Institutes for Research, 2007).

[52] Debra Viadero, "House Subcommittee Approves Proposal to Abolish NAEP Board," *Education Week*, February 9, 1994.

[53] Gerald Bracey, "Oh, Those NAEP Achievement Levels," National Association for Secondary School Principals, August 23, 2005,

[54] "Status of Achievement Levels," National Center for Education Statistics, U.S. Department of Education, available online at http://nces.ed.gov/nationsreportcard/nde/help/qs/Status_of_Achievement_Levels.asp.

[55] Richard M. Jaeger, "Reporting the Results of the National Assessment of Educational Progress."

[56] Diana Jean Schemo "Nation's Charter Schools Lagging Behind, U.S. Test Scores Reveal" *New York Times*, August 17, 2004.

[57] 2005 NAEP Results: State vs. Nation (Washington, DC: Achieve Inc., 2005), available online at www.achieve.org/node/482.

[58] "Mapping 2005 State Proficiency Standards Onto the NAEP Scales (NCES 2007-482)," National Center for Education Statistics, U.S. Department of Education.

[59] Answering the Question That Matters Most: Has Student Achievement Increased Since No Child Left Behind? (Washington, DC: Center on Education Policy, 2007).

[60] Lynn Olson, "Standards Get Boost on the Hill," *Education Week*, January 17, 2007.

[61] "Beyond NCLB: Fulfilling the Promise to Our Nation's Children," The Commission on No Child Left Behind, February 2007.