The NAEP Long-Term Trend Assessment: A Review of Its Transformation, Use, and Findings

Lawrence C. Stedman

March 2009

Paper Commissioned for the 20th Anniversary of the
National Assessment Governing Board
1988–2008

Lawrence C. Stedman is Associate Professor of Education at the State University of New York at Binghamton. He has used NAEP for many years in research on educational patterns and trends.

During the past 25 years, the country witnessed a dramatic transformation of the National Assessment of Educational Progress (NAEP). Actions by the Educational Testing Service (ETS), Congress, and the National Assessment Governing Board fundamentally changed NAEP's role in federal educational policy and the nation's schools. Developed in the 1960s through a privately funded initiative, NAEP began as a voluntary program run by a state consortium with financial support from the Department of Health, Education, and Welfare. It later became a congressionally legislated program administered by one of the country's premier testing organizations and overseen by a federally mandated public board.[1] Over time, NAEP's focus and scope changed substantially, expanding to grade and state testing, reporting by achievement levels, and, as part of No Child Left Behind (NCLB), requiring participation to receive Title 1 funds. NAEP was no longer a program whose results were reported in passing, but had become central to monitoring the nation's progress in achievement and equity. One major change was splitting NAEP into two *separate* programs: 1) the main assessment that tested students in grades 4, 8, and 12 in diverse subjects and 2) the long-term trend assessment that tracked performance in reading, writing, math, and science at ages 9, 13, and 17 as NAEP had done since 1969.

This paper describes how NAEP's trend assessment changed, its use in national educational discussions, and its major findings. From the earliest days, NAEP trends have figured prominently in debates over the decline of excellence, and extra attention is devoted to that issue. The paper also discusses the way in which NAEP trends have been used in evaluations of NCLB and the minority-majority achievement gap. A final section addresses the future of the long-term trend assessment. Its utility has been sufficiently questioned that the Board has considered eliminating it.

Origins and Transformation

While this paper focuses on the long-term trend assessment, by necessity, a discussion of its history must begin with the original NAEP program. For its first 20 years, there was only one program and a central part of its mission was to monitor trends. Its architects planned to regularly test nationally representative samples in 10 learning areas covering literacy, math and science, social studies, fine arts, and workplace skills. NAEP was the brainchild of three major figures of the era: Frances Keppel, U.S. Commissioner of Education; John Gardner, president of the Carnegie Foundation; and Ralph Tyler, one of the nation's foremost curriculum specialists and evaluators (Finder, 2004; Jones, 1996; Vinovskis, 1998).[2]

From the outset, NAEP was to be a different type of program that uses innovative formats, item reporting, and measurement of national achievement (Fitzharris, 1993; Jones, 1996). Its approach was greatly shaped by a Technical Advisory Committee (TAC) led by John Tukey, Princeton's famed mathematician. The TAC included Cronbach and Abelson (Jones, 1996, p. 21). Its successor, the Analysis Advisory Committee, included Glass, Moses, and Mosteller as well as Tukey. With such a stellar cast from the worlds of statistics and applied research, NAEP's design was elegant and visionary, a compelling departure from the standardized testing programs that had dominated most of the 20th century. Given that learning had many sources—the school, the media, reading, and home—the program would assess broad domains that went beyond school curricula. The tests would involve real-world materials and favor constructed-response items over multiple-choice items. To provide a more comprehensive

portrait of the nation's people, NAEP would assess young adults (ages 25–36) as well as those *outside of school* at lower ages.

The program was to be an evolving one, capturing trends over time as well as reflecting changing conceptions of what people should know. It would test a recurring set of items, but change them over time to stay current. Stability and change were viewed in tandem, not in opposition, as has been the case in recent years. As often happens as a program takes shape, it ends up differing in profound ways from what its designers intended. Hard realities set in, political interests become involved, and different philosophies take control.

Controversial Beginnings

Throughout its history, NAEP has been embroiled in controversy. When NAEP was proposed, many worried it would impose a national test, undermine local goals, and result in federal control of the schools (Finder, 2004; Vinovskis, 1998). The most pointed critique came from Harold Hand, a University of Illinois professor of education who had vigorously opposed the conservative assault on the public schools in the 1950s. He argued that NAEP would lead to a "centrally controlled curriculum" and its proposed school and district comparisons would "stultify the curriculum," undermine equal opportunity, and encourage cheating by "students and teachers alike" (Hand, 1965, p. 9). The criticisms met a receptive audience. Even those who helped create NAEP recognized it could unwittingly establish standards and lead to teaching to the test (Stake, 2007). They gave control of NAEP to the Education Commission of the States (ECS) instead of the federal government and dropped the idea of measuring the performance of individual schools, districts, and states. The reports would focus on national results, highlight performance on individual items instead of scale scores, and leave interpretations to outside panels of educators and citizens (Ahmann, 1979; NAEP, 1970).[3] As ECS stated in 1970, "the Commission does not want to assume the role of 'authority'" (Foreword, NAEP, 1970). The program's objectives would be shaped with broad public input. "Educators, scholars, and lay persons from all over the country" were drafted to determine what was important for young people to know (Foreword, NAEP, 1976).

Although this history has been largely forgotten, it remains relevant. Over the past generation, as NAEP adopted achievement standards and began comparing states and even districts, and as NCLB compelled participation, the early concerns seem prescient. Five phases of change affected NAEP: implementation in its first decade, a shift in control to ETS in the early 1980s, congressional actions in the late 1980s, Governing Board decisions in the 1990s, and NCLB legislation in 2002. The following sections describe the changes in each phase and how they affected the trend assessment.

Early Testing and Implementation

NAEP's first tests cover citizenship, writing, and science in 1969, followed by assessments in reading, literature, music, and functional literacy in 1970 and 1971.[4] From 1972 to 1974, NAEP also assessed social studies, mathematics, career and occupational development, and art. By 1983, NAEP reported, "All areas except career and occupational development have been periodically reassessed to detect any important changes" (NAEP, 1983, p. ix). During its

first decade, NAEP also assessed an impressive array of other areas, including health awareness, consumer skills, basic life skills, and energy awareness and attitudes. All told, 16 areas had been assessed.

The program was steadily realizing its ambitious goals. The early reports were rich ones, detailing item performance and national trends, and providing diverse interpretations by educators about their meaning for teachers and schooling. For each assessment, new test items were developed, while a core of common items was repeated to establish trends. As NAEP was implemented, however, several changes proved necessary; one of its architects called them "compromises" (Jones, 1996, p. 16). The objectives and test items became more closely aligned with school curricula; the results reported more often by percentages on sets of items than on individual ones; and its sampling, which had included young adults and out-of-school youth, was narrowed to those in school. While the tests remained groundbreaking, they also ended up being more traditional than had been originally planned. Still, the innovative philosophy that characterized the program in the mid-1960s lasted for nearly 20 years. However, in 1983, a fundamental shift occurred whose impact reverberates today.

The Shift to ETS: "A Grand Scheme"

In 1983, operational control of NAEP switched from the governors' Education Commission of the States (ECS) to ETS. Under ECS, the assessment process had been decentralized and diverse. In addition to TAC, different organizations, such as American Institutes for Research and Science Research Associates, constructed the tests and carried out the analyses (Jones, 1996, p. 16). The reports varied and the idiosyncrasies were charming and useful. Under ETS, the process was standardized. ETS revamped the test frameworks, applied item-response theory (IRT) to the results, put most scores on a 0–500 scale, and developed *performanc*e levels, which corresponded to the 150, 200, 250, 300, and 350 scale points. The skills and knowledge at each level were briefly described, along with a representative set of problems. All test takers were arrayed along the single scale.[5]

These were profound changes. NAEP no longer focused on item analysis and the percentage of correct answers. Trends were now tracked by scale scores and the percentage that achieve each level. The original plan of eschewing standards setting had been overturned. Other basic changes occurred. In some cases, testing by age was expanded to grades 3, 7, and 11, foreshadowing later grade level assessment. In others, the number of common items used to establish trends was reduced and fewer were released publicly. ETS introduced the idea of The Nation's Report Card, which helped capture media attention, and expanded into new areas. In the 1980s, NAEP added computer competence, U.S. history, geography, and document literacy and assessed the literacy of young adults. Over 20 areas had been assessed and trends were reported for many of these areas.

Many of the changes proved controversial. The shift to item-response theory and scale scores deeply troubled NAEP's original developers and other commentators. They were concerned that ETS had sacrificed a rich, multidimensional, item-based assessment for traditional norm-referenced testing. Tyler was "dismayed that the ETS had taken over the assessment" and "regarded that powerful organization as a redoubt of psychometricians

preoccupied with the bell curve" (Finder, 2004, p. 33). He complained that ETS had cast aside extensive work and the public shaping of objectives and instead, as he put it, "just took items off their shelf" (p. 33). Jones (1996, p. 17) felt that the tests "began to focus on desired curricula rather than on curricula already in place" and worried whether students had the "opportunity to learn answers to the questions asked." McLean and Goldstein (1988) decried ETS's scaling as "not connected to processes of teaching and learning" and labeled the reading scale "a fiction" lacking in validity. To its critics, the ETS standards were an example of, as Gene Glass once stated, a "grand scheme" erected on a "fundamental unsolved problem" (Glass, 2003, p. 1; Ho and Haertel, n.d.).

There was some truth to the critiques. ETS set the proficiency levels at five points along the bell curve. The 300 and 350 levels were one and two standard deviations above the mean, which meant only a small percentage would reach them. Over the years, critics have repeatedly assailed the school system because few students reach the top levels, yet this was psychometrically predetermined. IRT levels-based reports often produced harsher interpretations of national performance. In 1986, NAEP analysts looked at the percent correct on history items and reached a moderate conclusion that "the majority of high-school juniors do have some basic information about U.S. history" (Applebee et al., 1987, p. 10). For the 1988 history report, ETS analysts used proficiency levels and bluntly concluded, "Across the grades, most students have a limited grasp of U.S. history" (Hammack et al., 1990, p. 10). Still, scale scores remained a viable way of gauging achievement trends. Several studies indicated IRT modeling worked reasonably well (Pellegrino, Jones, and Mitchell, 1999, p. 70) and several NAEP reports showed trends in percentages correct and scale scores were similar (e.g., Dossey et al., 1988, p. 132). Another set of changes to the assessment was brewing, however.

Congressional Mandates

In 1988, Congress legislated key features of NAEP, including tracking achievement trends. The Hawkins-Stafford Elementary and Secondary School Improvement Amendments of 1988 established the Board and formalized NAEP's testing schedule. NAEP was expected to "report achievement data on a basis that ensures valid reliable trend reporting" and to

> collect and report data on a periodic basis, at least once every 2 years for reading and mathematics; at least once every 4 years for writing and science; and at least once every 6 years for history/geography and other subject areas selected by the Board. (p. 218)

Although NAEP had assessed over 20 areas, only 6 were mentioned. Priorities were established, but the Board still had the latitude to assess other areas and reassess areas that were previously tested. Congress also mandated that both ages (9, 13, and 17) *and* grades (4, 8, and 12) must be tested every 2 years. The age and grade testing begun by ETS was now formalized.

While Congress gave administrative control to the Commissioner of Education Statistics, it entrusted the Board with overseeing NAEP's methods, analysis, and reporting, and gave it responsibility for developing the test objectives for each assessment (Hawkins-Stafford, 1988, p. 220). It also charged the Board with two new missions: "developing standards and procedures for interstate, regional and national comparisons" and "identifying appropriate achievement

goals for each age and grade in each subject area to be tested" (p. 220). These actions would have serious repercussions, including changing the status of the trend assessment.

The Governing Board, 1988 to present

Under the Board, NAEP underwent basic changes. First, it was split into two distinct programs: (1) the main assessment, including state testing, focused on grades 4, 8, and 12; and (2) the long-term trend assessment of national performance at ages 9, 13, and 17. Second, NAEP's purposes expanded dramatically. It was now in the business of state comparisons. The Nation's Report Card had also become the states' report card. Some of NAEP's developers were troubled by this development. Tyler's biographer, Finder (2004), described his reaction:

> Tyler also regretted another distortion of NAEP's mission, which resulted from a congressional law of 1988—the introduction of competition. The law permitted statewide administration of NAEP exercises and, because statewide results are accessible, they encourage invidious comparisons between the states. (p. 34)

Third, with public participation, the Board established *achievement* levels (basic, proficient, and advanced) for *each* grade in the main assessment. These levels differ from the performance levels that had been set up by ETS in that they are "judgmental" rather than descriptive, yet are still overlaid on the 0–500 scale (Campbell, Hombo, and Mazzeo, 2000, p. 16). Both assessments reported national trends using scale scores. ETS or its chosen partners continued to conduct them.

The Board levels proved controversial. Over the past 20 years, a series of congressionally mandated evaluations by the General Accounting Office (now Government Accountability Office) and the national academies questioned their validity *and* utility (Stedman, 1998). Even the Board's own technical analysis found problems with how the levels were defined and applied (Chelimsky, 1992, p. 3). Since the mid-1990s, NAEP report cards have included disclaimers that the levels remain "developmental" and that "the process for setting them remains in transition" (O'Sullivan et al., 2003; Reese et al., 1997, p. ii). These evaluations were sharply criticized, however (Carroll, 1988; Reese et al., 1997; Center for Public Education, 2008). The NAEP report was found to be rife with "errors and inaccuracies" and to have used "professionally unaccepted standards of evidence" (Cizek, 1993, p. 4). Resolving such issues is beyond the scope of this paper, but it is worth noting that both the Board and the Commissioner of Education Statistics have repeatedly "affirmed the usefulness of these performance standards for understanding trends in achievement" (Perie, Grigg, and Dion, 2005, p. 2). Several evaluators also felt they could "be of use in describing changes in student performance over time" (Pellegrino, Jones, and Mitchell, 1999, p. 176).

Although unintended, the trend assessment was relegated to a secondary role and its scope was dramatically curtailed. What had begun as a program to track achievement in 10 areas, and then grew to over 20, was now reduced to only 4: reading, math, writing, and science.[6] In contrast, the main assessment was adding subjects, with regular testing in history, geography, civics, and the arts. After being conducted every other year from 1988 to 1996, the trend assessment was put on a longer time cycle. The main assessment was now conducted more often and, with its reporting of state results, generated more public attention. It also became a more

meaningful gauge of achievement. It was periodically updated to reflect changing curricula and new conceptions of learning, while the trend assessment remained fixed in frameworks developed by ETS in the 1980s. It also used more innovative formats and tested higher order skills more thoroughly.[7]

By 1999, the number of subjects in the trend assessment was further reduced. Although writing had been assessed in that year's study, technical difficulties precluded reporting the results. It was subsequently dropped entirely. Science was dropped after 1999 while awaiting a curricular and test overhaul—there was a clear need to modernize an assessment that had begun before DNA testing, widespread concerns over global warming, and the Hubble Space Telescope. Although the Board (2002) stated that science would eventually be put back into the trend assessment, the 2004 NAEP trend report noted, "According to NAGB's new policy...science and writing would be assessed only in main NAEP" (Perie, Moran, and Lutkus, 2005, p. 2). The NAEP testing schedule no longer lists science in the trend assessment (National Center for Education Statistics, 2008c).

The purposes of the two assessments came to be viewed differently. The authors of the 2004 NAEP trend report (Perie, Moran, and Lutkus, 2005, p. 2) described the difference this way:

> In this way, main NAEP can provide valid data for those seeking evidence for contemporary questions, and long-term trend NAEP can provide data for evaluating change over long periods. (p. 2)

Historically, though, only one assessment had been needed to fulfill both purposes. NAEP was the trend assessment. As Tukey and his colleagues noted in 1971, "National Assessment's main purpose is to measure change in what children and young adults know and can do" (Tukey et al., 1971, p. 1). In spite of the purported distinction, the main assessment also provided trend data, including changes in national scale scores over time and changes in the percentages at each achievement level. While its trends were not as long as those of the trend assessment, they still often spanned a decade or more. This further threatened the relevance of the trend assessment, but its status would become more secure with the passage of No Child Left Behind.

No Child Left Behind

NCLB completed the transformation of NAEP.[8] For the first time, states and districts were required to participate in NAEP's fourth and eighth grade reading and math state assessments if they accepted Title 1 funds. Public and private school achievement would be assessed "at least once every 2 years, in grades 4 and 8 in reading and mathematics" and at "regularly scheduled intervals" at grade 12 (p. 1898). After fulfilling the basic requirements, the act permitted additional, regular national assessments at grades 4, 8, and 12 in areas such as "writing, science, history, geography, civics, economics, foreign languages, and arts, and the trend assessment" (p. 1899).

Trend reporting was to be a feature of all NAEP assessments. NAEP should use timely reporting that "includes trend lines" and the Secretary of Education should ensure NAEP was "reporting the trends in academic achievement in a valid and reliable manner" (p. 1898). NCLB also explicitly mandated the long-term trend assessment. NAEP was to "continue to conduct the trend assessment of academic achievement at ages 9, 13, and 17 for the purpose of maintaining data on long-term trends in reading and mathematics" (p. 1899). The trend assessment in science and writing was not mandated. The main assessment would use Governing Board levels but, as a House-Senate conference report specified, the trend assessment would be conducted as it had been before NCLB (National Assessment Governing Board, 2002, p. 2). The long-term trend assessment, at least in reading and mathematics, was now a fixture of federal legislation.

The reach of the main assessment grew. The 1988 Hawkins-Stafford prohibition against assessing districts and schools was dropped in 1994 (Jones, 1996, p. 17). In 2002, NAEP began assessing urban districts (NCES, 2008b). In 2005, the Board studied, but rejected, the sharing of results with schools and students as a means of increasing 12th grade participation (Cavanagh, 2005; Fields, 2008). In 2008, NAEP was authorized to conduct a pilot assessment in 2009 that would permit state comparisons at 12th grade for the first time (Cavanagh, 2008). Such a sweeping expansion mirrors the fears expressed when NAEP was first proposed in the mid-1960s. What was supposed to be an unobtrusive assessment at the national level is now being used to compare states and districts. Required participation and state reports have greatly increased attention to NAEP. This is a concern. Teaching to the test can artificially inflate scores. In the 1970s, scores were higher on released NAEP items (Jones, 1996, footnote 6, p. 21). NAEP is in an educational catch-22. The more relevant it becomes and the more publicity it receives, the poorer a measure of trends it could be.

Over the past several decades, therefore, NAEP was greatly transformed. To its proponents, such changes represented an important modernization of NAEP. It now had technical sophistication, enhanced validity, consistency, and relevance to policymakers. By comparing states and districts, it could spur reform and improve education. To its critics, the changes represented an abandonment of fundamental principles and the creation of a testing enterprise that was doing more harm than good. NAEP had become intrusive and was emphasizing psychometric efficiency and uniformity over meaningfulness and improvement of student learning. I have reviewed the history here not to adjudicate these competing perspectives, but to illustrate the depth of the transformation. In spite of the profound changes, NAEP retained its twin purposes of measuring achievement and charting growth over time. Trends, however, are only as good as the instrument gathering the data. The next section discusses NAEP in comparison to other measures of long-term achievement trends.

The Mismeasure of Man

Besides NAEP, four other indicators have been widely used to judge long-term achievement trends: college admissions tests (especially the SAT), commercial standardized tests, state trends, and then-and-now-studies. Each has major limitations (Stedman, 1998, 2003).

The SAT

The SAT illustrates the problems of using admissions tests to establish achievement trends. It is taken by an elite group of college-bound students (mostly those with high grade point averages, high class rank, and strong academic transcripts) and not the typical high school student. Its scores do not even represent college-bound seniors well because many take other admissions tests such as the ACT. In the past, observers often misread SAT trends as evidence of a major educational decline, yet ignored an enormous compositional change in test takers that accounted for much of the change. The College Board has warned against using the SAT to measure national educational quality. As an aptitude test, its relevance to school curricula is questionable. For most of the post-World War II period, the SAT consisted of sentence completions, verbal analogies, reading passages, and general math problems. The SAT also emphasizes rapid response over thoughtfulness, as students have to answer questions at the rate of more than 1 question a minute (over 200 problems in 3 hours). This is a poor way of assessing the overall impact of schooling.

Commercial Standardized Testing

Commercial standardized test score data are often cited; however, they are a notoriously unreliable indicator of national achievement trends. Standardized tests are constructed in such a way that small changes in performance can produce large changes in grade equivalents and percentiles. On the SRA test, for example, 12th graders dropped a full grade level in reading during the 1970s, but this was only from 72 to 68 percent correct (a four-point drop). Trends were derived by comparing performance on new and old versions of the tests, but publishers' equating studies usually did not involve nationally representative samples. Instead, only a few school districts or a fraction of the national sample took both tests. Sometimes only parts of the tests were given. Trends were also confounded by teaching to the test and familiarity with repeatedly administered tests. Even test publishers cautioned against using equating data to gauge national trends.

State Trends

Several commentators used state trends to argue that national achievement rose from the 1950s to the mid-1960s before declining greatly in the 1970s. The state data were too sketchy, however, to have produced any firm conclusions about *national* trends (Stedman and Kaestle, 1987). One reviewer, for example, cited data from *only three* states, and his Idaho and New Hampshire data came from only one grade and his West Virginia data were mostly from a different period. Iowa was often cited, but it is predominately rural with few minority students. Trying to establish national trends by amassing state data is problematic because the tests and scores were not comparable. State trend data were also undermined by the problems plaguing commercial tests—the changing relevance of test content and the effects of test familiarity. Like the SAT, unaccounted for changes in the composition of test takers were a major problem.

Then-and-Now Studies: Generational Changes

In another approach, researchers repeat a test given years or decades earlier. Such then-and-now studies produced dramatically conflicting claims. Several educators used them to assert that achievement improved during the 20th century before collapsing in the 1970s. This "first major skills decline in American educational history" (Copperman, 1978, p. 39) was supposedly so severe that students had fallen well behind earlier generations. In contrast, others argued that 1970s students were doing as well as those of the 1940s and 1950s. Some even claimed that "almost all the results" showed improvement and the current generation was doing better (Berliner and Biddle, 1996; Bracey, 1992; National Council of Teachers of English, 1996). On their face, however, the studies provided little evidence for such sweeping claims (Stedman, 1996a, 1996b, 1998). The historical record showed a mixed pattern: some up, some down, with many roughly level. The best estimate was that students of similar background performed at about the same level throughout the 20th century (Stedman and Kaestle, 1987; Stedman, 2003). In general, the studies were plagued by serious problems. Many involved only small, local samples and were not nationally representative. For any given period, the few studies available were too geographically scattered to determine national trends. Researchers typically failed to assess the huge changes in student composition that had occurred in the decades between tests. They also failed to account for several major factors that disadvantaged the contemporary students: outdated test material, younger ages, and lower dropout rates.

Why NAEP?

In contrast to these other measures, NAEP provides the most credible gauge of trends (Stedman, 1998). It tests nationally representative samples every few years in the major academic areas, is well grounded in school content, and employs diverse testing methods. The test frameworks were developed with wide public input and reflect a broad consensus about what students should know and be able to do. NAEP has long used modern conceptions of learning and measurement. From the outset, its tests have included authentic materials and complex tasks. The reading tests ask students to respond to stories, poetry, articles, and advertisements. The math tests require students to interpret graphs, plot data, use calculators, and read tables. The geography and history tests have used a rich panoply of maps, photographs, cartoons, and magazine covers. In the past decade, the tests in the main assessment have focused even more on open-ended items. In geography, over 50 percent of the testing time was devoted to constructed-response items (Weiss et al., 2002, p. 6). Such rich assessment instruments have provided compelling information about student performance and achievement trends.

How NAEP Trend Data Have Been Used

Given its striking qualities, it is little wonder that NAEP has been at the heart of debates over school quality. In each decade since NAEP began, a major national report has featured its trend data. This section reviews the trend data and discusses NAEP's role in the great achievement debate of the 1990s.

On Further Examination

In the 1970s, the College Board panel examining the SAT decline noted that NAEP results "differed considerably from those on most other standardized examinations" (Advisory Panel, 1977, p. 23). The Panel found there had been only slight declines in reading and writing, while functional literacy and basic reading skills had improved. NAEP was in its infancy, however, and the Panel looked forward to future results. It welcomed NAEP's planned testing of adults 26–35 and those *out of school* at age 17 as it would provide needed information about those age groups. It noted that only the Armed Forces Qualifications Test (AFQT) provided data about the *entire* age group, not just students or the college bound, and that AFQT scores had been rising (p. 23).[9] The Panel speculated that while students might be "less well equipped for what college has traditionally required," "the general ability level of youth as a whole increased" (p. 24).

A Nation at Risk

In 1983, *A Nation at Risk* cited NAEP data as support for its claims that student performance was deficient (National Commission on Excellence in Education, 1983). Its data handling, however, provided better evidence of how data are misused than of a major decline in the nation's schools (Stedman and Smith, 1983). The Commission claimed that 13 percent of 17-year-olds were functionally illiterate, yet that figure was inflated. Students had to score above 75 percent to be considered literate. Had researchers chosen the 60 percent level, only about 3 percent would have been labeled illiterate. The report misconstrued literacy in black-and-white terms. Many so-called "functional illiterates" could handle many tasks. Worse, the Commission ignored gains during the 1970s when NAEP repeated the test two more times.

The Commission cited a "steady decline in science achievement scores of U.S. 17-year-olds as measured by the NAEP" (NCEE, 1983, p. 9), yet the decline had been small (less than 5 percentage points). The report failed to mention that high school reading and writing were generally steady during the 1970s, while younger students' reading and math scores were improving.

The Commission also noted that many 17-year-olds could not carry out "higher order" tasks. While correct, it neglected to mention the NAEP percentage had remained steady. Within a year after *A Nation at Risk,* scores on the Iowa Tests of Basic Skills reached their highest level. Instead of a "rising tide of mediocrity," the report could have proclaimed a "rising tide of test scores" (Stedman and Kaestle, 1985).

National Education Goals Reports

In the 1990s, the National Education Goals Panel (NEGP) used data from the main NAEP assessment to monitor progress as part of Goals 2000. Goal 3 was that *all* 4th, 8th, and 12th graders would be competent in challenging subject matter, defined as the Governing Board proficient level (NEGP, 1995, p. 36). To track trends in achievement and in closing racial, ethnic, and gender gaps, NEGP used the percentage of students achieving proficiency or better.

NAEP's main assessment proved so useful in revealing trends that it suggested the long-term trend assessment had become superfluous.

Over time, NEGP shifted from being a *national* goals panel to being a crusader for state reform. Its "national" reports became dominated by *state* comparisons and celebrated which states were the "highest performing" and which the "most-improved" (NEGP, 1999b, p. 31). The Panel even issued a states-only report highlighting NAEP reading scores (NEGP, 1999a). The language shifted as well: the phrase "national progress on core indicators" became the "U.S. Scorecard" (cf. NEGP, 1995, p. 7 and NEGP, 1999b, p. 16). What NAEP's early critics, and even some of its advocates, had feared had come to pass—a competitive emphasis on state performance rather than a sensible and nuanced use of national NAEP data. The entire effort had become simplistic, focused on scorekeeping, and detached from the real world of schooling and learning. Curiously, the Panel did *not* issue a culminating report in 2000 evaluating how well the national goals had been achieved. The 1999 NAEP trend report could have played a central role in such an evaluation.

The Great Achievement Debate

From the mid-1970s through the 1990s, NAEP trend data were at the heart of a fierce debate over the state of U.S. schooling (Stedman, 1998). On one side, a diverse set of educators and school critics claimed there had been a major decline in achievement. In such provocatively titled books as *Dumbing Down Our Kids* and *The Decline of Intelligence in America,* they blamed a lowering of academic standards and argued that the nation's well-being was at stake. Their case was mired in hyperbole reminiscent of the 1950s conservative attacks on progressive education in such works as *Quackery in the Schools*. Phrases such as "massive decline" and "unremitting fall" recurred frequently. *A Nation at Risk* decried a "rising tide of mediocrity that threatens our very future as a Nation and a people" (NCEE, 1983, p.5). In *The Literacy Hoax,* Copperman (1979, pp. 48, 101) cited NAEP trend data as proof of a "sharp drop-off" in science, "devastating declines" in civics, and a "deterioration of writing skills." In *Cultural Literacy,* Hirsch (1987, p. 7) argued NAEP data provided compelling "evidence for the decline in shared knowledge." Sykes (1995) echoed such claims but, unlike some who felt a recovery was under way due to the standards movement (Ravitch, 1995), he used NAEP data to argue reform efforts were failing.[10]

> More than a decade after A Nation at Risk drew attention to the nation's educational mediocrity, the reading proficiency of nine- and thirteen-year-olds has declined even further. (p. 20)

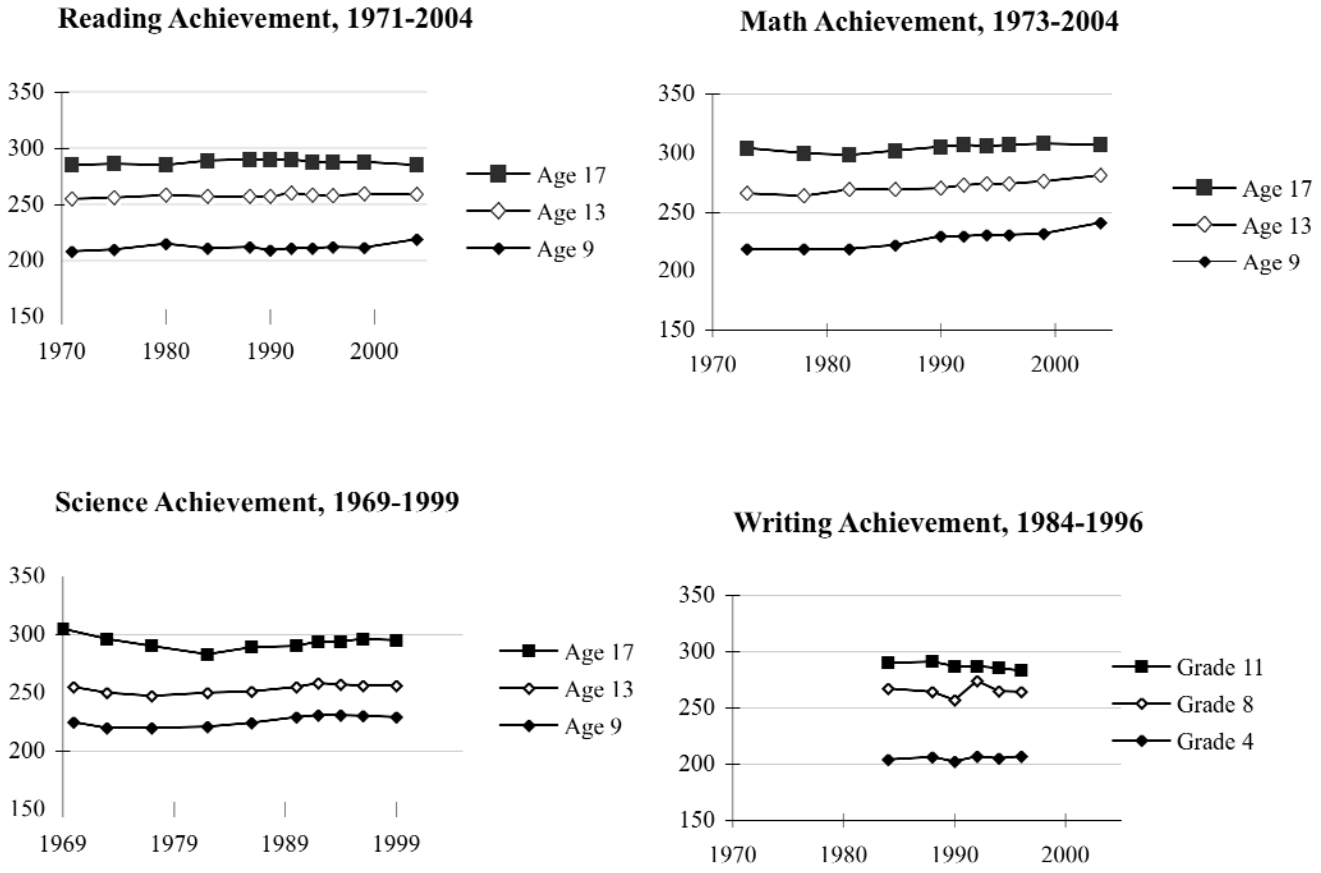
In response came equally sweeping claims by those who argued public schools were doing "better than ever" (Bracey, 1992, p. 107) and NAEP scores had reached "all-time" highs (Bracey, 1995). Principals, the NCTE, and op-ed writers echoed these sentiments (Stedman, 1995). Some revisionists brashly asserted that the SAT was the only test that ever suggested a decline and argued that the educational crisis had been manufactured by right-wing forces (Berliner and Biddle, 1995). Such pointedly titled works as *The Manufactured Crisis* and *The Way We Were?: Myths and Realities of America's Student Achievement* presented this dissenting viewpoint. Such charged rhetoric suggested the trend evidence had not been carefully analyzed

by either side (Stedman, 1996b, 1998, 2003). The next section shows that this proved to be the case.

Findings: The Light at the End of the Tunnel

NAEP's long-term trend assessment has tracked performance in reading, mathematics, science, and writing. What do the data show? The patterns are striking (figure 1). With some fluctuations, scores have changed little in over a generation. The overall impression is one of stable achievement (2008 data are expected shortly).

Figure 1



Reading and math data are from Perie, Moran, and Lutkus, 2005, figures 2-1 and 2-4, pp. 10, 17. Science data are from Campbell, Hombo, and Mazzeo, 2000, figure 1.1, p. 9. Writing data are from Campbell, Voelkl, and Donahue, 1998, figure 1, p. vi. Data for these graphs are in the appendix. Note: The NAEP scale runs from 0 to 500, yet 150 to 350 is a fairer representation (2 standard deviations on both sides of the mean). The full scale is unrealistic (no age group scores below 150 or above 350), while letting software set the scale is problematic as it uses only a small portion of the scale, just enough to span the data. That fills a graph with the data, thus

exaggerating small fluctuations. NAEP reports typically truncate the scales—in the 1990s, they often used 170 to 320; most recently 200 to 320.

The one noticeable decline was among 17-year-olds in science in the 1970s. Scores dropped 20 scale points; however, in percentage terms, the decline seems more modest—6½ points in 13 years.[11] Seventeen-year-olds had a minor decline in math (6 scale and 3 percentage points). During the 1980s, they recovered the ground they lost and are now doing as well in both areas as they had in the early 1970s. Younger students improved in math and science, but the math upswings were sudden and chaotic. Most of the gain by 9-year-olds occurred in just two assessments.

The early writing assessments were not subject to the scale because they used a different framework. Still, a series of NAEP reports reveal that writing, like reading, has remained roughly stable since 1969 (Applebee et al., 1990; Stedman, 1998). Literacy is central to schoolwork so its constancy contradicts assertions about a major decline. In general, achievement in 2004 is similar to what it was in the early 1970s. Given that, it is hard to argue that the current generation of students (or that of the 1990s) is doing substantially worse—or better—than their parents did. One looks in vain for an educational collapse or dramatic gains in the graphs; there is no rising tide of mediocrity or wholesale improvement due to the standards movement.[12]

Civics

Although it was not part of the separate trend assessment, NAEP has assessed civics several times. Given our Jeffersonian ideal of education as civic preparation and the sweeping claims of sharp, even "devastating" declines in civics, the data should be closely examined. Determining trends is not easy as the assessments shifted from ages to grades, from percent correct to scale scores, and used different scales. Putting them together, however, reveals a rough pattern of a modest decline over several decades followed by level performance (table 1). Both 13- and 17-year-olds' civics scores declined several percentage points in the early 1970s. Thereafter, fluctuations were small. In 1998, NAEP found civics achievement comparable to that of 1988 (Weiss et al., 1998). There were only small changes (2 percentage points) at the lower grades and "no significant difference" for 12th graders (Weiss et al., 1998, p. 24). In 2006, the main assessment showed that civics achievement had been steady since 1998 at grades 8 and 12. Over the past 20 years, therefore, civics achievement has changed little. As in other areas, stability had taken hold.

Table 1

Civics Achievement, Age 17 and Grade 12

Civics area	1969	1972	1976	1982	1988	1998	2006
Age 17							
Citizenship Knowledge (percent correct)	73		65*				
Social Studies Knowledge (percent correct)		64	59				
Civics Proficiency (scale score 0 to 100)			61.7	61.3	59.6*		
Grade 12							
Civics Performance (percent correct)					68	66	
Civics Proficiency (scale score 0 to 300)						150	151

Civics Achievement, Age 13 and Grade 8

Civics area	1969	1972	1976	1982	1988	1998	2006
Age 13							
Citizenship Knowledge (percent correct)	65		62*				
Social Studies Knowledge (percent correct)		50	48				
Civics Proficiency (scale score 0 to 100)			49.1	49.1	50		
Grade 8							
Civics Performance (percent correct)					64	62*	
Civics Proficiency (scale score 0 to 300)						150	150

*A statistically significant change at the .05 level. Other changes were not significant.

The data for 1969–1976 are from NAEP, 1978, p. 69. Data for 1976–1988 are from Anderson et al., 1990, p. 13. Data for 1988–1998 are from Weiss et al., 2001, p. 9. Data for 1998–2006 are from Lutkus and Weiss, 2006, p. 7.

Contrasts With the Results on Other Measures

In general, NAEP showed steadier scores and smaller declines than those on many standardized tests in the 1970s. Some commentators speculated that NAEP tests involved lower level skills and were easier, which meant that students could perform better. In fact, the percentage of questions answered correctly by 17-year-olds on NAEP was comparable to that on other tests, and a similar proportion of the tests was devoted to inferential skills (Stedman and Kaestle, 1986, appendix). Scores on some commercial tests were also steady during the period.

The contrast between generally level NAEP scores and the large apparent SAT decline was not surprising. The drop on the SAT was relatively small in actual performance terms and, as noted, was largely caused by compositional changes in test takers. Whatever skills the SAT measured did not decline between the 1950s and the 1980s. We know this because nationally

representative samples of high school students were tested on the PSAT (a short version of the SAT) five times between 1955 and 1983. Although there were some fluctuations, performance in the early 1980s matched that of the early 1960s (Stedman, 1998). The Advisory Panel (1977, pp. 22–23) also found that, in spite of their lower SAT scores, students were now scoring higher on College Board achievement exams in English, the sciences, and foreign languages.

Appraising Trends

The judgments of level trends often fell along ideological lines. Some argued that flat scores and rapidly growing expenditures showed the system was inefficient, while others felt maintaining performance in the face of social upheavals demonstrated that schools had an unrecognized resilience (Stedman, 1998). It is striking that achievement remained steady even as school populations had become more diverse socioeconomically. In the 1970s and 1980s, child poverty rates rose, the percentage of single-parent households almost doubled, and it was more likely that both parents or a student's sole guardian would be working outside the home. Other changes, though, should have improved performance: parents were better educated, drug use dropped greatly, reported school safety improved, and pupil-teacher ratios dropped. Evaluating the impact of these changes depended more on the eye of the beholder than on systematic analysis. As we saw in the great achievement debate of the 1990s, careful, balanced treatments of NAEP's trend data have not been the norm. This was particularly true when the 2004 NAEP trend report came out. After a generation of reform efforts and a major federal push with NCLB, strong judgments were to be expected.

Reactions to the 2004 Trend Report

The 2004 trend report generated a parade of upbeat, promotional press releases. Even though most scores had changed little and gains were modest, the U.S. Department of Education and leading educational organizations trumpeted the success of school reform. Secretary of Education Spellings boldly asserted the 2004 study was "proof that *No Child Left Behind* is working—it is helping to raise the achievement of young students of every race and from every type of family background" (U.S. Department of Education, 2005). The American Federation of Teachers (AFT) argued the improvement was under way well *before* NCLB and was due to its hard-working members, its reading program, and the standards movement (AFT, 2005). The National Council of Teachers of Mathematics (NCTM) attributed the "clear improvement" to the "continuing effect of mathematics standards" (NCTM, 2005).

All-time high claims, which had been part of a "new mythology" about U.S. achievement (Stedman, 1995), returned with vigor.[13] NCTM's president claimed "our elementary and middle school students are performing core skills like computation better than at any time in the thirty years since the test was first administered" (NCTM, 2005). NCES (2005) emphasized that reading and math scores of 9- and 13-year-olds were higher in 2004 than when NAEP began. As in the 1990s, such claims were sometimes true in a strictly numerical sense, but overall they painted a misleading picture. Even at younger ages, reading scores were only a few points higher in 2004 than in 1971 and, as the graphs showed, stable scores would have been a fairer description.[14] The case in math was a stronger one, but the jumps at the younger ages were anomalous. The focus on younger students disregarded high school achievement, which was

certainly *not* at all-time highs.[15] Over 30 years, high school students' scores had changed little. In 2004, their reading score was exactly the same as it had been in 1971, while their math score was the same as in 1992 and only 3 points higher than it was in 1973 (Perie, Moran, and Lutkus, 2005, pp. 10, 17).

The impact of national standards on achievement was questionable. Younger students' math performance was the most touted indicator, but showed little improvement in the decade after NCTM released its standards. In the 1990s, scores for 13-year-olds went up only 6 points, while those for 9-year-olds rose only 2 (Campbell, Hombo, and Mazzeo, 2000, p. 9).[16] Whatever gains students made vanished by high school. The official pronouncements of "success" had given short shrift to stagnating high school scores. Still, the flurry of publicity showed the trend assessment continued to play a central role in national discussions.

A Verdict on No Child Left Behind?

The Secretary of Education's statement that the 2004 NAEP trend findings were proof NCLB was succeeding rapidly came under fire. Smith (2007) pointed out that NCLB was too new for it to have much impact on the 2004 results. It had passed in 2002 and was not fully implemented until 2005. He used state and national NAEP data to reach devastating conclusions:

> The best evidence available suggests that NCLB may actually be reducing student gains in reading and making no difference in math achievement. (p. 1)
>
> These findings not only contradict Secretary Spellings' contention that NCLB is working, but they also suggest that all the 1999–2004 gains that she touted came in the first three years of the period, before NCLB was enacted. (p. 4)

Smith also summarized 13 studies showing that NCLB accountability features were "minimal, implementation erratic and weak, and evidence of effectiveness largely lacking" (p. 4). He pointedly questioned the congressional rush to renew NCLB.

NAEP data were also used as evidence that NCLB was improving literacy and helping English-language learners, but such claims were disputed (Institute for Language and Education Policy, 2007; Krashen, 2006; Shanahan and Hynd-Shanahan, 2006). In addition, NCLB supporters cited increasing percentages achieving proficiency in *state* testing programs, but these involved limited or short-term data, and state results were often inflated compared to NAEP's results (Hoff, 2007).

Formal evaluations of NCLB's impact support Smith's conclusions. Fuller et al. (2007) examined data from the NAEP trend assessment, state testing programs, and main NAEP. They reported that youngest students gained the most, but that reading "growth flattened out in fourth grade over the 3 years after enactment of NCLB" (p. 275). Fourth grade math progress was "more buoyant" (p. 275), but also slowed *after* NCLB. In a study for Harvard's Civil Rights Project, Lee (2006) used trend data from the main NAEP assessment and also concluded that NCLB had not improved achievement. Still, both the claims of success and these findings were a bit premature in reaching a judgment about the efficacy of NCLB. As Smith (2007) pointed out,

> In fairness, it should be noted that a considerable body of research suggests that large scale education reforms rarely show effects within three years. So while the reading tests seem to be saying that something changed around 2002 to arrest U.S. reading gains, it could be argued that, at this point, we really can't say with confidence whether NCLB is working to improve, maintain or limit student literacy. We only know that no plausible case can be made that it is working. (p. 4)

We should await, therefore, the findings of the 2008 NAEP trend study before reaching any definitive judgments about NCLB. I predict, though, that they will not resolve the debate over NCLB's impact. As happened with the 2004 study, NCLB supporters will celebrate even minor gains as a resounding success. If the scores are level or even if there are some declines (which is unlikely given all the testing that is going on), they will call for a redoubling of the standards-testing effort, especially at the high school level. This was already foreshadowed by Secretary Spellings' reaction to the 2004 trend report (U.S. Department of Education, 2005):

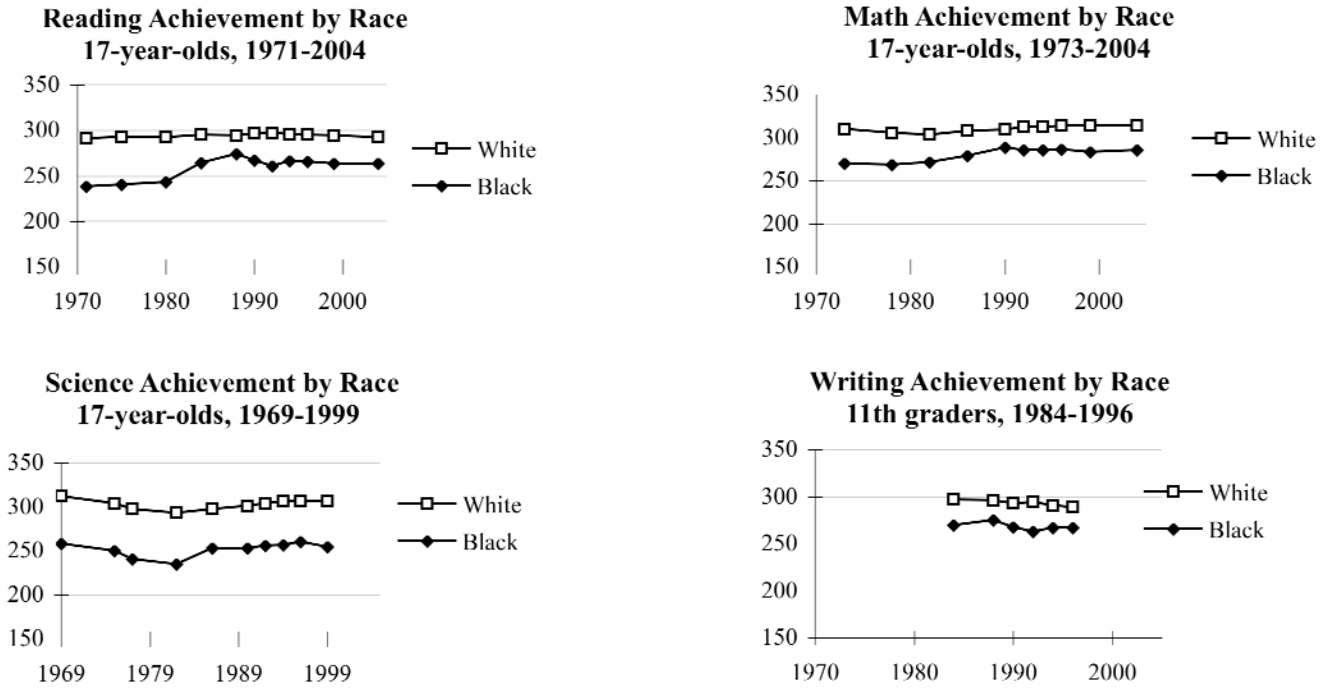
> So I am pleased with today's results, but in no way completely satisfied. We are at the beginning of the journey and certainly have room for improvement, particularly at the high school level. We must support older students with the same can-do attitude that helped their younger brothers and sisters.

The 2008 trend findings—no matter what they are—will not persuade NCLB detractors to change their minds either. If scores continue to be sluggish, they will tout that widely as evidence of its failure. If, as I suspect, scores improve somewhat at the younger ages or if gains have accelerated post-NCLB, they will likely attribute that to rampant testing and teaching to the test rather than to genuine improvement in learning. As in the great debate over the test score decline, judging NCLB is now more a matter of self-interest and ideology than a sober assessment of the evidence. The findings pertaining to equality of opportunity were treated similarly.

Equality of Opportunity

Over the past 15 years, concerns grew over the minority achievement gap and educators worked hard to reduce it. NCLB required scores to be reported by race and ethnicity. The 2004 trend report led to proclamations the gap was closing. The AFT highlighted the "narrowing achievement gap" (AFT, 2005) while Secretary Spellings asserted that the "achievement gap that has persisted for decades in the younger years between minorities and whites has shrunk to its smallest size in history" (U.S. Department of Education, 2005). In touting NCLB, President Bush (2008) noted, "Scores for minority and poorer students are reaching all-time highs in a number of areas, and the achievement gap is closing." In fact, while the achievement of younger black and Latino students improved somewhat, they still lagged well behind and the gaps were similar to or larger than they had been in the late 1980s. Using NAEP trend data, Lee (2006) determined the gaps had not narrowed after NCLB. The chasms in science were even larger and growing and commentators had overlooked the persistent gaps among high school students. The black-white reading gap among 17-year-olds in 2004 was *larger* than it had been in the late 1980s, while those in math and science were somewhat greater than they had been in the early 1990s (figure 2). The writing gap fluctuated but was about the same in 1996 as in 1988.

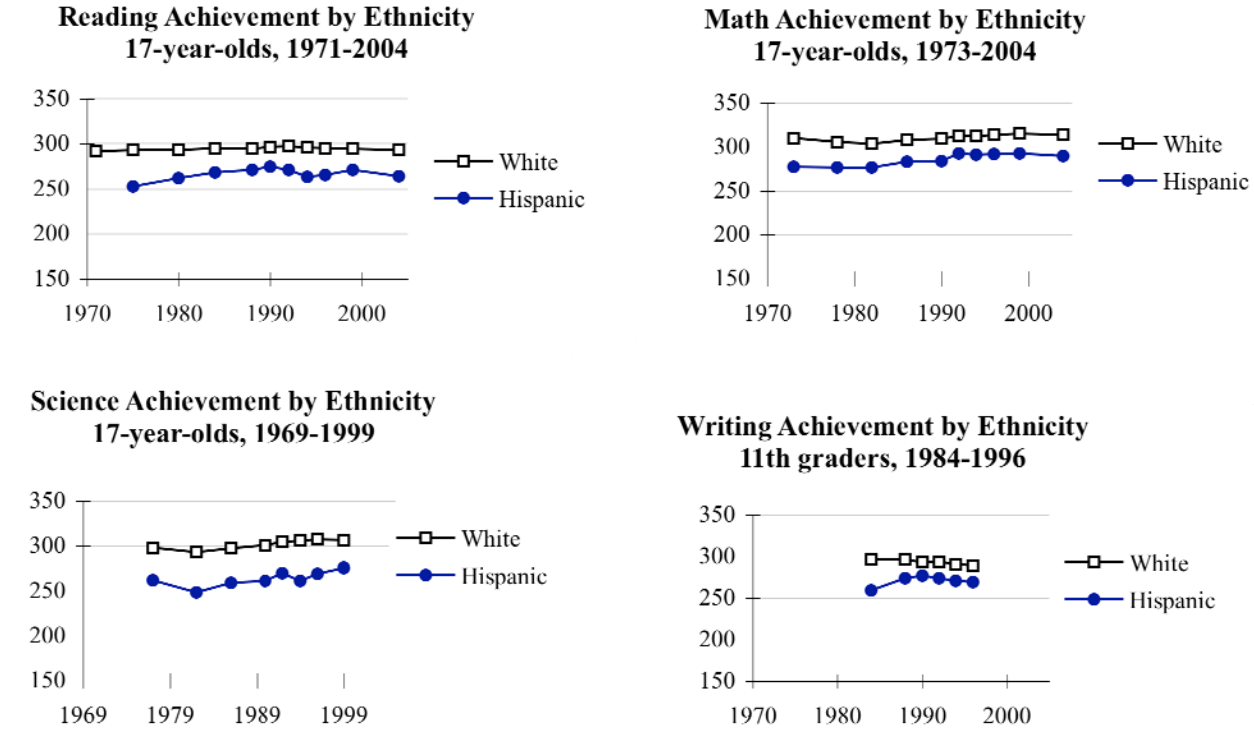
Figure 2



Reading and math data are from NCES, 2006, tables 110 and 121, pp. 175, 189. Science and writing data are from NCES, 2003, tables 127 and 116, pp. 158, 148.

The graphs reveal an unexpected pattern. Scores for minority high school students generally improved during the Reagan years and leveled off during the Clinton years. In reading, the black-white gap was cut by over 60 percent in the 1980s, but then widened and is now nearly 50 percent higher. There also has been little closing of the gaps in math and science during the past decade. In science, the black-white gap remains almost as large as it had been in 1969. The patterns and gaps were similar for Hispanic high school students (figure 3).

Figure 3



Disaggregation of data for Hispanic students began later than for African American students. Reading and math data are from NCES, 2006, tables 110 and 121, pp. 175, 189. Science and writing data are from NCES, 2003, tables 127 and 116, pp. 158, 148.

> Orfield (2006) reached a sobering conclusion:
>
> It is important to keep in mind that the NAEP does show substantial declines in racial achievement gaps in the 1970s and early 1980s, when more of the civil rights and anti-poverty efforts of earlier reforms were still in operation. The strict standards-based reform effort that swept the county after the 1983 *A Nation at Risk* report has not shown similar benefits on achievement gaps. (p. 6)

Such persistent achievement gaps raise questions about the impact of the standards movement and NCLB. They should have been a cause for concern and action, not proclamations of success. NAEP trend data have shed light in three other major areas: whether there had been a decline in high achievers, the extent of private-public school differences, and gender inequality.

Decline at the Top?

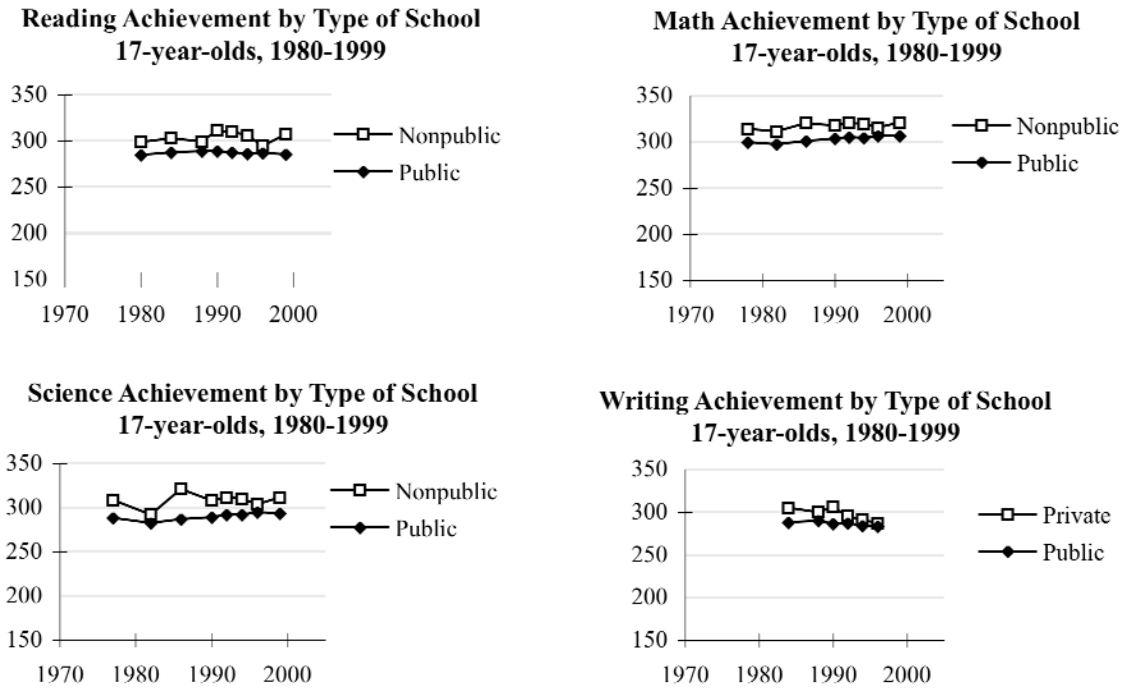
Part of the concern over the decline in excellence has centered on a loss in high-scoring students. This arose relative to college admissions exams, but the evidence was mixed and the causes complex (Stedman, 1998). The ACTs showed little change in top scorers and the fact that elite colleges dropped the SAT helped reduce its high scorers. In NAEP, the proportion of 17-

year-olds reaching the highest performance level (350) has been roughly stable for several decades (Campbell, Voelkl, and Donahue, 1998, p. xii; Campbell, Hombo, and Mazzeo, 2000, p. 25; Perie, Moran, and Lutkus, 2005, pp. 15, 23). The proportion reaching the next highest level (300) shows a mixed pattern. In math and science, it initially rose, but has changed little since the early 1990s. In reading, it has been dropping for over a decade and is now back to the level of the early 1970s. In writing, it declined during the 1980s and 1990s.

Private vs. Public Schools

NAEP data also have been used to compare private and public schools (Stedman, 1996a). Overall, the differences have been relatively constant and not substantial, averaging around 11–18 scale score points (figure 4) (calculated from Campbell, Hombo, and Mazzeo, 2000, pp. 53–55). Much of it can be accounted for by differences in socioeconomic status. Private school families are better educated, have larger incomes, and more often live in upscale communities.

Figure 4



"Nonpublic" and "private" designations are those of the original data source. Nonpublic school results were not reported in 2004 because participation rates were too low. Reading, math, and science data are from Campbell, Hombo, and Mazzeo, 2000, tables B.20 to B.22, pp. 119–121. Writing data are from NCES, 2003, table 116, p. 148.

The extreme fluctuations in the gaps raise reliability questions. In reading, a sudden drop in private school achievement in 1996 reduced the gap to a fraction of what it was before and after that date. In science, a small gap suddenly widened, tripling in 1986 due to a jump in private school scores, but then closed again in the 1990s. Such erratic patterns are likely related

to sampling fluctuations. In the 2004 trend report, private schools' nonparticipation rates were so large that their results were not even included (Perie, Moran, and Lutkus, 2005, p. 27).

An AFT analysis of NAEP data in the early 1990s showed that public school students who had taken advanced math courses overcame the usual gap and slightly outperformed private school students (Shanker, 1991). From the data, some inferred that the private-public difference was a matter of curriculum (Berliner and Biddle, 1996), but the public school students taking college-prep math were an elite group that should have scored higher. Shanker (1991) also reached a conclusion that received insufficient attention, namely, that students in both sectors were performing poorly.

Gender Gaps in Achievement

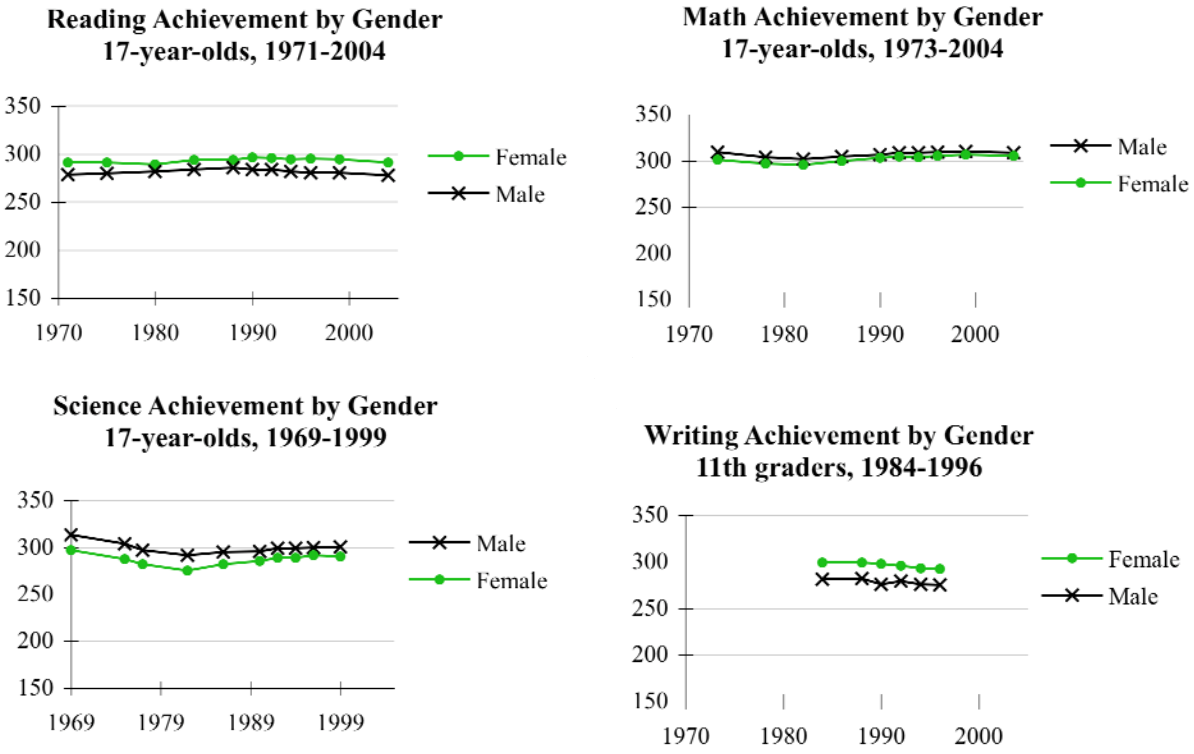
In the 1990s, NAEP data also showed up in debates over how the nation's schools were treating girls. In *Failing at Fairness,* Sadker and Sadker (1994) noted that NAEP

> displays a familiar school picture, one with a commonly accepted gender divide: boys overtaking girls in math and increasing their superiority in science, and girls maintaining an advantage in reading and writing. (p. 138)

College admissions tests, however, "the most important tests," "paint a much more depressing portrait" (p. 138). In *Who Stole Feminism?,* Sommers (2004) countered that the gender gap was overblown as NAEP showed "the math and science test differentials are small compared to large differentials favoring girls in reading and writing" (p. 160). This debate echoes in the present, with recent assertions that it is boys, not girls, who are in trouble.

What do NAEP trends tell us about gender gaps? Over the past three decades, high school gender gaps have been small—much smaller than gaps by race and ethnicity—and the achievement trends for boys and girls have been similar (figure 5).

Figure 5



Reading and math data are from NCES, 2006, tables 110 and 121, pp. 175, 189. Science and writing data are from NCES, 2003, tables 127 and 116, pp. 158, 148.

The gender gap in reading was the narrowest during the Reagan years and then nearly doubled, mostly due to slippage in boys' scores. Still, the gap remains small and, in 2004, both genders scored about where they had in the early 1970s. Contrary to stereotypes, the gender gaps in math have been tiny and the smallest of the four areas. The gender gap was largest in writing, with a female advantage that changed little in the 1980s and 1990s.

Low Levels of Achievement

A focus on trends has obscured the persistent problem of low achievement. Our concern should not only be whether students are doing better than before but whether they are doing well enough to participate in the democracy, meet the demands of today's society, and fulfill their potential. NAEP's findings are telling (Stedman, 1998, 2003). On average, less than half of our 17-year-olds reach the 300 level and very few (2 to 10 percent) reach the 350 level (table 2); yet, these are the levels purportedly needed for work, higher education, business, and government (Applebee, Langer, and Mullis, 1989, pp. 22–23).[17]

Table 2

17-year-olds' Achievement on 1996–2004 Trend Assessments

Subject	Year	300 level+	350 level+
Reading	2004	38%	6%
Math	2004	59%	7%
Science	1999	47%	10%
Writing	1996	31%	2%
Average		44%	6%

Reading and math data are from Perie, Moran, and Lutkus, 2005, pp. 15, 23. Science data are from Campbell, Hombo, and Mazzeo, 2000, p. 25. Writing data are from Campbell, Voelkl, and Donahue, 1998, p. xii.

Performance has also fallen short in the main assessment. While some trends are up, most high school seniors are still not at or above the proficient level and large percentages (on average a third) fail to reach the basic level.[18] In reading, the percentage achieving proficiency has been declining and is now down to 35 percent. Math performance is poorer; only about a quarter are at or above the proficient level. History performance is the worst; more than half the students fall below the basic level.

In spite of the concerns over the validity of the levels, the performances are troubling. Many of the items that define the upper levels on the long-term trend assessment are simple ones. The 350 level in math includes "routine problems involving fractions and percents" (Dossey et al., 1988, p. 31), while that in history includes items asking only for recognition of basic information (Hammack et al., 1990, pp. 25–26). Although the level-setting procedure *initially* limited the proportion of students at the top, it did not prevent that proportion from growing. Unfortunately, as noted, the percentage of high school students reaching the highest level has changed little in over three decades.

Examinations of NAEP test items avoid the level problems and reveal deep and persistent problems in each subject (Stedman, 2003). Our high school students continue to struggle with middle school math, including percents, finding area, estimations, and simple algebra (NAEP, 2008).[19] They have not learned basic information from their U.S. history classes, such as the purpose of the Monroe doctrine or the fact that the Scopes trial dealt with evolution (Stedman, 2003; NAEP, 2008). In the mid-1990s, more than half were not familiar with the Camp David accords or realize that U.S. foreign policy after World War II was dominated by the effort to contain communism. Geography knowledge has also been limited. In the 1980s and 1990s, almost two-thirds of high school students could not locate Southeast Asia, the Persian Gulf, or Saudi Arabia on a world map. This is disturbing given the country's history in the Vietnam War and ongoing problems in the Middle East.

Perspectives on Low Achievement

In spite of these findings, one's judgments should be tempered. High school students have done well in some areas. They know elementary math operations and can read simple graphs. They are aware of some leading documents from U.S. history and such major countries as Canada and Germany. Most can handle basic writing mechanics. We should remember that many high school students have lacked opportunities to learn this material in other ways—by traveling, visiting museums, discussing it with parents, and taking college courses. Still, the material is basic (much from seventh and eighth grades) and is routinely taught in school, so one would expect high school students to have learned it.

Motivation to do well may be a factor (Stedman, 1998). NAEP tests 12th graders in the spring when they have less academic focus. If NAEP tests were high stakes, scores would likely be higher, but performance has not improved on NAEP items embedded into state testing programs. NAEP's tests are much shorter than other standardized tests (only 45–50 minutes long), and the reduced burden should enable students to do better.

Overall, students may be doing worse than suspected. NAEP no longer tests dropouts (a sizable 25 percent of the student body), who would probably score lower.[20] To reach the 300 or 350 levels, students only have to correctly answer 65–80 percent of their problems (Mullis et al., 1991, p. 218). A higher standard would mean even fewer students have reached these levels. Many tests, such as those in science, have relied on multiple-choice items on which students do better. As NAEP includes more constructed-response items to better gauge learning, more deficiencies will appear. The preponderance of evidence indicates students are struggling with basic curricular material and have major deficiencies in their understanding. The observation of a NAEP report on its first 20 years still applies: "very few students demonstrate that they can use their minds well" (Mullis, Owen, and Phillips, 1990, p. 10).

Non-achievement Findings

The trend assessment has surveyed students about their schooling and extracurricular activities, including math instruction, science equipment and computer use, television viewing, homework time, and reading for fun. Some contextual factors are less informative. Reports of course-taking, while interesting, are not a substitute for NAEP's high school transcript studies. Several useful items are no longer reported: what students read; reading activities related to friends, bookstores, and libraries; family television rules; types of math instruction; and attitudes toward mathematics. This has broken trend lines and indicates that tracking test score trends has mattered more than understanding their context and how it has changed over time.

Classroom Instruction

Math has been dominated by traditional, teacher-centered instruction. High school students watch their teachers work problems, listen to them explain lessons, and take tests (table 3). There has been a marked increase in all these activities, especially testing. At the same time, a growing majority reports that discussions take place in their math classes. Still, most students do

not have an opportunity to work problems at the board and very few complete reports or math projects.

Table 3

17-year-olds' Report of Mathematics Classroom Activities (percentages)

Often	1978	1996
Watch the teacher work mathematics problems on the board?	80	87
Listen to a teacher explain a mathematics lesson?	79	86
Take mathematics tests?	64	84
Discuss mathematics in class?	51	62
Work mathematics problems on the board?	28	27
Make reports or do projects on mathematics?	2	5

Data are from Campbell, Voelkl, and Donahue, 1998, p. 88, table 4-4.

Literary Activity

Over the past generation, educators have grown increasingly concerned about a decline in reading and have vigorously debated how reading is taught (Chall, 1995; Flesch, 1981; Goodman, 2004; Kaestle et al., 1991; Taylor, 1998; Thimmesch, 1984). Several literacy experts argued that an abandonment of phonics had caused a decline in reading achievement while others argued that, by neglecting real literature, the basic-skills approach had contributed to aliteracy—people who know how to read but do not. During the past 5 years, the National Endowment for the Arts (NEA) has released three provocative reports on the issue. Its 2004 report, *Reading at Risk,* revealed major declines in adults' leisure reading of "novels, short stories, plays, or poetry" between 1982 and 2002 (NEA, 2004, p. ix). In its 2007 report, *To Read or Not To Read: A Question of National Consequence,* NEA compiled the results of different assessments, including NAEP's trend studies. NEA's chairman reported, "The story the data tell is simple, consistent, and alarming" (NEA, 2007, p. 5). Both literary and nonliterary reading had declined, especially among college graduates. The decline in reading was linked to worsening reading achievement. Leading literacy specialists, however, such as Krashen and Shanahan, felt that the evidence about a reading decline was unclear and that the reports were not nuanced enough (Rich, 2007, 2008).

The most recent NEA report, *Reading on the Rise,* indicates there has been a modest increase in fiction reading in the last few years (mostly among young adults), yet the overall rate of reading any type of book outside of school and work has continued to drop (NEA, 2009; Thompson, 2009).

The NAEP trend data speak to this issue and indicate there is growing aliteracy. In 2004, fully a third of 17-year-olds reported that they rarely, if ever, read for pleasure (table 4). The percentage had nearly doubled from 20 years earlier. Students also read much less often as they grow older (table 5). In 2004, over half of 9-year-olds reported reading for fun almost daily

compared to only a fifth of 17-year-olds. That gap has been widening over the decades. Still, it should be noted that most students did some reading for pleasure, with about half reporting they read regularly, daily, or weekly (Perie, Moran, and Lutkus, 2005, p. 55).

Table 4

17-year-olds' Reported Frequency of Reading for Fun (percentages)

	1984	1994	1999	2004
A few times a year, or never or hardly ever	19	24	28	33

Data for 1984, 1999, and 2004 are from Perie, Moran, and Lutkus, 2005, p. 55. Data for 1994 are from Campbell et al., 1996, p. 152.

Table 5

Percentage Reporting Reading for Fun "almost every day"
(percentages)

	1984	1994	1999	2004
9-year-olds	53	58	54	54
13-year-olds	35	32	28	30
17-year-olds	31	30	25	22

Data for 1984, 1999, and 2004 are from Perie, Moran, and Lutkus, 2005, p. 55. Data for 1994 are from Campbell et al., 1996, p. 152.

Homework

There continues to be a pitched battle over homework. Concerns have been raised about whether there is too much or too little, whether it is busywork or meaningful, and if its purpose is to instill a work ethic or improve learning. Arguments against excessive homework have been laid out in *The Homework Myth* (Kohn, 2006) and *The Case Against Homework* (Bennett and Kalish, 2006). In contrast, other educators blamed the test score decline on a lack of homework. *A Nation at Risk* argued that American students spend "much less time on school work," used time "ineffectively," and that schools needed to assign "far more homework" (NCEE, 1983, pp. 18, 25). There have been conflicting research reports about homework time and its utility (Stedman, 1998). The international assessments found great variations in homework but with inconsistent, often negligible, effects on achievement (Stedman and Smith, 1983; Stedman, 1997). (There is an irony here. If homework time goes up, there is less time for reading for pleasure.)

The trend assessment provides 30 years of data pertinent to this debate. It has repeatedly asked students "How much homework did you do yesterday?" In 2004, only a third of high

school students reported doing an hour or more the day before (table 6). In the decades after *A Nation at Risk* came out, time spent on homework declined. By 2004, over a quarter reported "no homework was assigned." The combined percentage that did not do homework or did not receive any grew in the 1990s and is now nearly 40 percent. It exceeds the percentage doing 1 hour or more daily.

Table 6

17-year-olds' Report of Time Spent the Day Before on Homework
(percentages)

	1980	1984	1994	1999	2004
1 hour or more	33	40	39	35	33
Did not have	32	22	23	26	26
Had but did not do it	12	11	11	13	13

Data are from Perie, Moran, and Lutkus, 2005, p. 51.

While these data seem alarming, there are caveats. It is important to remember that these are self-reports. The little or no homework percentage may be artificially high because students were surveyed in the spring. For many 17-year-olds, it was their last semester and may not reflect what they generally experienced in high school. Other surveys showed increases. A 1980–2002 national comparison found that sophomores' weekly homework time had increased dramatically (Cahalan et al., 2006; NCES, 2007).[21] The contradictory results may be partly due to the different groups surveyed. Recollecting a week's homework time, however, is less accurate than reporting the previous day's homework. Other NAEP trend evidence supports a more sanguine view of academic involvement. Since 1984, the amount of academic-related reading that high school students report has not declined (table 7). Definitive answers will require that NAEP gather corroborating evidence from schools, teachers, and parents.

Table 7

17-year-olds' Report of the Number of Pages Read Daily in School and for Homework (percentages)

	1984	1994	1999	2004
16 or more pages	35	36	36	38
5 or fewer	21	21	23	21

Data for 1984, 1999, and 2004 data are from Perie, Moran, and Lutkus, 2005, pp. 52, 53. Data for 1994 are from Campbell et al., 1996, p. xxiii.

Conclusion: Shining Star or Faded Glory?

For 40 years, the trend assessment has been central to discussions of U.S. education. Even so, the Board has considered eliminating it or merging it with the main assessment. In 1996, in a unanimously adopted policy statement, the Board called for a transition to a single program. Board members felt it was becoming "impractical and unnecessary to operate two separate assessment programs" and that the main assessment could "become the primary way to measure trends in reading, writing, mathematics, and science" (National Assessment Governing Board, 1996, p. 10). They noted the "tension between stable measures of student achievement and changing curricula" would be a "continuing policy matter," but felt bridge studies could maintain historical time series. A few years later, an evaluation committee of the National Research Council (NRC) also recommended merging the two assessments. In *Grading the Nation's Report Card,* it observed that the "proliferation of multiple independent data collections—national NAEP, state NAEP, and trend NAEP—is confusing, burdensome, and inefficient" (Pellegrino, Jones, and Mitchell, 1999, p. 56).[22]

By 2002, however, NCLB mandated that NCES retain a separate trend assessment and the Board reconfirmed its commitment to the program. The main assessment would be "designed to reflect current curriculum content in the nation's schools" while the trend assessment would "reflect a fixed curriculum framework" (National Assessment Governing Board, 2002, p. 3). As we shall see, this was an artificial distinction and poses vexing issues. The following paragraphs examine the leading arguments (both pro and con) about keeping the long-term trend assessment. Notwithstanding NCLB's mandate, it is worth examining these arguments because a repeal of NCLB, or a change in this provision, could mean that the Board will have to decide whether or not to continue it.

Consistent Test Items. The strongest argument for retaining the long-term trend assessment is that it has tested the same items for more than three decades and so has a unique ability to track changes. It is supposedly based on a single curricular framework and unchanging tests first administered in 1969 or the early 1970s. This is a widely held belief. Secretary Spellings described the assessment as "using the same exact test in reading and mathematics for over 30 years" (U.S. Department of Education, 2005). NCTM (2004) noted, "The same test items have been used for mathematics since 1973." Even NCES (2008d) perpetuates this view, stating that content "has remained essentially unchanged since first administration (1971 for reading, 1973 for mathematics), although some changes were initiated in 2004."

The problem is that the trends have not been based on fixed tests, a fixed set of items, or even a fixed framework across the years (see appendix for details). In several cases, the same items were used in only two successive assessments. In others, they were repeated for several, but then the basket of items and even the frameworks were revamped. Often a subset of items was used to establish trends. The number of items varied greatly, as did the mix of constructed and multiple-choice items. The trend lines were often established by scaling scores, making extrapolations, and conducting bridge studies across different tests. There were other problems as well. Instead of being based on frameworks from 1969 or the early 1970s, the trend assessment has been anchored in curricular and testing frameworks established in the *mid-1980s*.[23] Trends

from 1969–2004 thus reflect how well students performed on *1980s* test material. That epoch's students had an advantage and the program *estimated* trends before 1977 in math and science.

This shows that both assessments establish trends across tests of varying content and frameworks. Still, the main assessment's trend lines have been shorter, and for the past 20 years, the trend assessment has used a basic framework and similar tests.

Outdated Frameworks. The NRC committee noted NAEP trend anomalies had "led measurement specialists to conclude that if you want to measure change, don't change the measure" (Pellegrino, Jones, and Mitchell, 1999, p. 78). Yet, it argued this would have

> some drawbacks over longer periods of time. It is not inconceivable that, held constant for long periods of time, frameworks become increasingly irrelevant by failing to reflect changes in curricula and instructional practice. (p. 78)

Long time horizons, such as 20–30 years, also prove problematic because student populations and social conditions change greatly. Trends have been mistakenly attributed to changes in school quality without considering changing contexts. This problem has plagued interpretations of then-and-now studies, the SAT decline, and NAEP trends. An evolving assessment that keeps up with the times, yet bridges to the past, is a better approach. This has now been formally instituted for the trend assessment (National Assessment Governing Board, 2002). In 2004, the decades-old curricular and testing frameworks were overhauled to better reflect today's schooling. Bridge studies have been used to link current and past results. While these actions are welcome, they mean the trend assessment no longer differs fundamentally from the main one, which also uses bridge studies to maintain trends.

Inconsistent Findings. Some arguments for merging the assessments have been weak. The two programs have sewn confusion by sometimes reporting different trends. The NRC committee was concerned that 8 of 12 overlapping periods had distinctly different trends (Pellegrino, Jones, and Mitchell, 1999, pp. 73, 77). Yet, this hardly justifies dropping the trend assessment. While it would end the confusion, such discrepant results raise questions about the reliability of *both* assessments and should be investigated, not buried.

Cost and Testing Burden. The NRC committee also felt that running multiple programs raises costs, increases complexity, and reduces participation in small states where schools are repeatedly sampled (Pellegrino, Jones, and Mitchell, 1999, p. 78). Yet, this argument more strongly supports eliminating the state assessment, which is far more expensive, complex, frequent, and intrusive. The burden is even greater now that NCLB requires state participation at fourth and eighth grade.

Lower-level Material. Another reason for discarding the trend assessment is that it supposedly tests lower order skills and relies excessively on traditional formats. The math test, for example, was described as having a "computational focus" in the 2004 trend report (Perie, Moran, and Lutkus, 2005, p. 4) and relying on multiple-choice questions with only "a few" short-answer questions (NCES, 2008d). Yet, the same test had been described in the 1990 trend report as covering a "range of tasks," including "multi-step problem solving and reasoning,"

"ability to read charts and graphs," and covering geometry, algebra, "linear equations, functions, and coordinate systems" (Mullis et al., 1991, p. 58). For much of the 1990s, the test included 56 open-ended questions (Mullis et al., 1991, p. 204)—hardly a "few"—while in 1999, more than a fifth of the items used to derive trends involved constructed responses (Campbell, Hombo, and Mazzeo, 2000, p. 85).

Nevertheless, there is some merit in the critique. In general, the main assessment uses more open-ended items and its questions are based on modern conceptions of learning and curriculum. Historically, for example, the trend science test relied *entirely* on multiple-choice items (Campbell, Hombo, and Mazzeo, 2000, p. 84), whereas the main science test involves constructed-response questions and even has some students conduct "actual experiments" and "record their observations and conclusions" (Grigg, Lauko, and Brockway, 2006, p. 3).

Redundancy. The most basic reason for dropping the trend assessment is that it has become redundant. The main assessment already generates trends spanning substantial periods. The 2007 math and reading main assessments traced performance back to 1990 and 1992 (Lee, Grigg, and Dion, 2007, pp. 24–25; Lee, Grigg, and Donahue, 2007, pp. 26–27). The 2006 history assessment traced achievement to 1994 (Lee and Weiss, 2007, p. 9). In some cases, though, the main assessment dropped trend coverage after adopting a new curricular framework (Grigg, Donahue, and Dion, 2007, p. 14). This happened with the 2005 12th grade math assessment. In such cases, bridge studies could maintain historical trends. As part of the NRC's work evaluating NAEP, Kolen discussed five different ways of achieving statistically well-grounded long-term trends using the main assessment (Pellegrino, Jones, and Mitchell, 1999, p. 79).

Similar Relevance. The NRC committee also reported a study comparing test items from the two assessments. In spite of being grounded in frameworks from different eras, middle-school math and science teachers and subject specialists judged the items equally relevant to contemporary curricula and concluded that students had similar opportunities to learn them (Pellegrino, Jones, and Mitchell, 1999, p. 78). If their content is similar, then it does not seem necessary to have separate assessments.

Limited Curricular Focus. The trend assessment is now down to only two subjects. Although reading and math are important, this is a thin reed upon which to judge school quality and long-term achievement trends. The main assessment provides richer information on more of the curriculum. It has tracked trends in science, writing, civics, geography, and history as well as math and reading. Economics was assessed in 2006; this subject and the arts are scheduled for repeated administrations (NCES, 2008c). Foreign languages are also being added.

Recommendations for the Future

Overall, therefore, there is a compelling case for dropping the trend assessment and relying on the main assessment to generate trends. The dual-testing system is expensive and confusing. The main assessment already provides useful long-term trends. It provides greater coverage of the curriculum, more authentic testing, and is better grounded in contemporary pedagogy. Still, no matter how NAEP gathers trend data, several reforms would improve its value to practitioners, policymakers, and the public (see appendix). Two reforms are considered

here; the first is the need for a deeper conception of academic preparation. "Reading achievement," for example, should encompass knowledge, understanding, and appreciation of literature. *Fundamentally, we care that students are well read, not just that they read well.* Are students familiar with leading authors and their works? Do they welcome the challenge, pleasure, and intellectual development that come from reading *books?* NAEP should routinely assess such matters. Without them, our measure of reading achievement is a hollow one. The country will continue trying to raise scores while neglecting the equally important issues of *how much* students are reading, the *quality* of what they read, and how much they are getting out of it. Skilled, engaged literacy also means writing well. The issue is not simply whether students can respond to NAEP's arbitrary prompts, but whether they can write with understanding about local civic issues; produce compelling essays, stories, and blogs; and write clearly in their history, science, and English classes.

A few times in the past, NAEP assessed reading and writing more richly. In 1970 and 1980, NAEP conducted *literature* assessments that explored students' written reactions to literature and the interconnections among reading, writing, and thinking (NAEP, 1982). In the late 1980s, NAEP assessed knowledge of literary works and discovered that high school students were unfamiliar with major women and African-American writers and classics by Shakespeare, Conrad, and Whitman (Applebee, Langer, and Mullis, 1988). In the 1990s, NAEP collected portfolios from English classrooms and found that even our students' best writing—done with consultation, revision, and choice of topics—was not that good (Gentile, Martin-Rehrmann, and Kennedy, 1995, pp. 31, 49). Only 4 to 12 percent of the eighth graders' papers achieved high marks. Such assessments should be repeated to establish trends and understand student literacy more fully.

Similarly, it is not enough for students to receive high scores on a math test; in addition, we want them to be comfortable with math and readily use it in the real world. We want our students to read history, visit historical sites and museums, and follow science developments with interest. Are schools stirring lasting intellectual passions? Are students expanding their horizons? Are they engaged in the world and drawing well on their academic pursuits? How are those things changing over time? Are the schools doing a better or worse job of inspiring students and unleashing their minds? Adding such measures and tracking their change would greatly enhance NAEP's value.

Assessing students' preparation for a democracy is also crucial and means going beyond measuring civics and history knowledge and gauging their active participation in politics, knowledge of contemporary affairs, and judgment of diverse news and information sources. To what extent is the Constitution a *living* document in students' lives? That is a vital question in an era when constitutional liberties are being eroded in the name of national security. As global warming transforms the planet, we need a scientifically literate population who will confront the self-serving arguments of economic interests and fashion creative solutions.

Schools exist for more than academics and raising test scores, and academics gain much of their value by being relevant to these other areas. A high-quality national assessment of educational progress needs to determine if actual learning is taking place, if tangible intellectual energy exists, if there is a commitment to social justice and planetary stewardship, and if the

foundation of a vibrant democracy is evident. Only then will the trend assessment we would like to have and the improvement in test scores that we hope to see (and that was overemphasized in the past) be truly meaningful.

Appendix

Establishing Trends

Contrary to a general impression, NAEP trend lines do not reflect a single testing framework and a repetition of the same test in every assessment. The variations in the assessments for 17-year-olds may be the best way to illustrate these changes.

Reading. The 1990s trend reports explained there were "some changes from test administration to test administration" in the reading assessment during the 1970s. The 1970–1975 reading trends for 17-year-olds were originally based on 85 items. In 1980, the reading assessment dropped 14 items because of their bias and stereotyping (errata sheet, NAEP, 1981) and recalculated trends on the remaining 71 items. In 1984, the trend assessment was revamped with a new testing design, a different type of matrix sampling, and a new reading framework. Reading trends in 1980–1984 were based on 53 common items (ETS, 1985, p. 29), although the 1984 assessment had many more items. In 1988, more changes were made and Westat, the NAEP data collector for ETS, conducted bridge studies to link the assessments. The 1984–1988 trends were based on 87 common items, a different subset of the 1984 assessment than had been used for the 1980–1984 trend line (Mullis and Jenkins, 1990, p. 48).

Although the 1994 trend study authors reported that reading passages and questions had been "kept essentially constant since 1984" (Campbell et al., 1994, p. 266), the reading trend test in the 1990s had grown to 112 items involving 34 reading passages and 9 constructed-response items. The 1999 trend study reiterated that the passages and questions had been "kept essentially the same since 1984" (Campbell, Hombo, and Mazzeo, 1999, p. 85), yet the reading test had been reduced to 95 items (of which 8 were constructed) and there were now 36 passages. Clearly, the assessments had changed since 1984 and were different from those of the 1970s.

Math. The math trend assessment used 61 items to compare the 1973, 1978, and 1982 performances, but used 383 items to compare those of 1978 and 1982 (NAEP, 1983, p. xiv). In 1986, it based trends on 94 of the items (Dossey et al., 1988, p. 125). In the 1990s, trends were based on 287 items (of which 56 were open-ended) (Campbell et al., 1994, p. 265). They were a subset of the items used in the full 1986 assessment and they used the same procedures from 1973. In 1999, the basis of the trends from 1986–1999 changed to 132 items (of which 29 were constructed) (Campbell, Hombo, and Mazzeo, 1999, p. 85).

Science. The first science trend assessment in 1969 used 124 exercises (Tukey et al., 1971, p. 6). The 1986 assessment used 111 items (Mullis and Jenkins, 1988, p. 137), but trends thereafter were established using a subset of 82 multiple-choice items (Campbell, Hombo, and Mazzeo, 2000, p. 84; Mullis et al., 1991, p. 204).

Writing. The writing assessments of 1974 to 1984 differed substantially from those used from 1984 to 1996 and were never placed on the same scale. The idea that NAEP establishes trends through a fixed, unchanging assessment is most applicable to the writing assessments of the late 1980s and 1990s. "There have been six national assessments of writing conducted during the school years ending in 1984, 1988, 1990, 1992, 1994, and 1996. The 1996 assessment

included the same set of 12 writing tasks that had been administered in the five previous assessments. Each of these trend assessments was administrated to nationally representative samples of students in grades 4, 8, and 11" (Ballator, Farnum, and Kaplan, 1999, p. 1).

Recommendations for the Trend Assessment

1. Reassess and establish trends in orphaned areas, especially in literature, energy attitudes and awareness, and health.

Given national debates about energy policy and concerns about obesity, AIDS, and STDs, this reassessment is vital. Academic aspects could also be explored in terms of scientific understanding of global warming, biological knowledge of diet and exercise, and understanding of food and energy politics.

2. Expand curriculum coverage, especially at the high school level.

Two subjects are too few for meaningful trends. Even the main assessment is missing key high school areas—English literature, math courses such as geometry and advanced algebra, and science courses such as biology and chemistry. Until it develops its own instruments, the Board and NCES could contract for national administrations of the College Board achievement tests (now known as SAT Subject Tests).

3. Assess other fundamental purposes of schooling.

This would include assessing students' preparation for democracy, ecological responsibility, and multicultural understanding. This means gauging involvement in community affairs, knowledge of global warming, and multicultural understanding. In the 1980s, NAEP assessed students' literature knowledge and included works by leading African American and women writers. Such an assessment would be well worth repeating. How well are students learning different languages and interacting with members of different cultures and nationalities? How much have students learned about racism and historical patterns of oppression? These are important matters in an era of globalization, religious and cultural conflicts, and for a country whose minority population is becoming a majority. Assessing such matters would help ensure a vibrant, working, multicultural democracy.

4. Gather more information about classroom practices, school conditions, and the nature of the curriculum.

Stultifying and psychologically damaging classrooms can raise scores. To what extent are teaching and curricula being narrowed, and understanding and creative projects being sacrificed? How great today are the savage inequalities that Kozol so eloquently documented? How much impact do they have on achievement? Without such information, achievement levels and trends will continue to be misinterpreted. NCES and the Board should also commission regular nationwide *qualitative* investigations of classroom life and school conditions.

5. Embrace a deeper conception of academic preparation.

We need to ensure that students have a deep understanding and appreciation of what they are learning, are engaged with it, and can apply it well—not merely that they can pick the right answer or write a brief response on a NAEP assessment.

Data for Figure 1

Reading Achievement, 1971–2004

Year	Age 17	Age 13	Age 9
1971	285	255	208
1975	286	256	210
1980	285	258	215
1984	289	257	211
1988	290	257	212
1990	290	257	209
1992	290	260	211
1994	288	258	211
1996	288	258	212
1999	288	259	212
2004	285	259	219

Math Achievement, 1973–2004

Year	Age 17	Age 13	Age 9
--	--	--	--
1973	304	266	219
1978	300	264	219
1982	298	269	219
1986	302	269	222
1990	305	270	230
1992	307	273	230
1994	306	274	231
1996	307	274	231
1999	308	276	232
2004	307	281	241

Science Achievement, 1969–1999

Year	Age 17	Age 13	Age 9
1970	305*	255	225
1973	296	250	220
1977	290	247	220
1982	283	250	221
1986	289	251	224
1990	290	255	229
1992	294	258	231
1994	294	257	231
1996	296	256	230
1999	295	256	229

*1969

Writing Achievement, 1984–1996

Year	Grade 11	Grade 8	Grade 4
1984	290	267	204
1988	291	264	206
1990	287	257	202
1992	287	274	207
1994	285	265	205
1996	283	264	207

Endnotes

1. Much thanks to Ray Fields, Assistant Director for Policy and Research for the Board, who provided helpful comments on the original manuscript. As Fields noted, mandated participation applies only to fourth and eighth grade state testing in reading and mathematics and ETS is no longer the sole or lead organization involved in the conduct of NAEP, as it was in the 1980s. See NAEP (2009) for a discussion of the history of NAEP contractors.

2. At the time NAEP was being developed, Tyler was the founding director of Stanford University's Center for Advanced Study in the Behavioral Sciences. He had written the influential 1949 volume, *Basic Principles of Curriculum and Instruction.* In the 1930s, he had been the research director for the Progressive Education Association's 8-year Study, a longitudinal study focused on the relationship between high schools and college that tracked students during their years in experimental high schools and through college (Finder, 2004; Fitzharris, 1993). The first report on the study (Aikin, 1940) is available at http://www.8yearstudy.org/index.html. A bibliography that lists the other volumes can be found at http://education.stateuniversity.com/pages/1947/Eight-Year-Study.html. There was also a 1938 book written by students at one of the high schools, *Were We Guinea Pigs?* The study is still relevant today and is one that educators, evaluators, and policymakers should be familiar with. See Cremin (1961), Kliebard (1995), and Krug (1972) for discussions of the history of the study and its findings. An ERIC or Google search will turn up many contemporary assessments of its significance and meaning.

3. In 1970, ECS explained that it would "issue National Assessment reports from time to time without interpreting the results or explaining their implications" (Foreword, NAEP, 1970). Instead, ECS routinely asked subject matter professional associations (NCTE, NCTM, NCSS, and NSTA) to independently write interpretive commentaries (Ahmann, 1979). Early NAEP reports were often written by panelists of leading math, social studies, or literacy educators and included quotes and observations from them about the diverse meanings of the findings (see, e.g., NAEP, 1976; NAEP, 1981, pp. 47–48). Many perspectives were heard. This approach changed dramatically under ETS.

4. The recurring chronological error in Governing Board and NAEP reports, state education department web pages, and news accounts is that the trend assessments began in 1970.

5. In a few cases, a different scale was used. Civics was first scaled from 0 to 100, which sowed confusion because many readers and even NCES itself sometimes thought ETS was reporting percentages rather than scale scores (see NCES, 1993, p. 124). Later, it was scaled from 0 to 300.

6. Many NAEP-related documents have implied the trend assessment always covered only four areas: reading, writing, math, and science. Adding to the confusion, NCES's chronology of NAEP assessments puts some, but not all, of the early assessments in the *long-term trend assessment* column even though many were repeated and achievement trends were established (NCES, 2008a).

7. These changes were prompted by requirements in the Board's test frameworks and specifications.

8. The full text of NCLB can be found at http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf. The NAEP sections begin on page 217. The key NAEP provisions of NCLB can be found at http://www.ed.gov/policy/elsec/leg/esea02/pg97.html#sec411 and at the Governing Board web site, http://www.nagb.org/about/plaw.html. A clear overview of the NAEP portion of NCLB, in nonlegal language, can be found at the NCES web site, http://nces.ed.gov/nationsreportcard/nclb.asp.

9. Fields (2008) rightly questions whether women were included. It appears that they were not included and the Advisory Panel was overstating its point. Still, the AFQT results indicate that scores for men, in *and* out of school, were rising during this period.

10. Several researchers presented more nuanced treatments of the NAEP trend and acknowledged where there had been improvements (Ravitch, 1995; Stedman, 1998; Steinberg, Brown, and Dornbusch, 1996). Still, Steinberg, Brown, and Dornbusch (1996, p. 45) argued that "the evidence clearly shows that the achievement drop is genuine, substantial, and pervasive across ethnic, socioeconomic, and age groups" and that "in many respects, student achievement is significantly lower than it was twenty-five years ago."

11. The science trend provides the best evidence for the formulation that there was a decline followed by a recovery due to the reimposition of standards. The trends in science from 1969 to 1977 and in math from 1973 to 1978, however, were actually *extrapolations* and appear with dotted lines in the graphs in NAEP trend reports. NAEP trend reports have cautioned about interpreting trends in those years (Campbell, Hombo, and Mazzeo, 1999, p. 93).

12. To be sure, there have been some impressive gains on the *main* assessment, though only in math and only at the *younger* ages. From 1990 to 2007, the percentage of fourth graders achieving proficiency in math on the *main* assessment tripled, while that for eighth graders more than doubled (from 15 to 32 percent) (Lee, Grigg, and Dion, 2007, pp. 8, 9, 24, 25). The annual growth rates, however, were more modest: only 1 to 1½ points per year. At those rates, it will take the schools several generations, from 40 to nearly 70 years, to bring all students to proficiency, the objective of Goals 2000 and NCLB. It will take nearly 150 years before even half reach the advanced level!

13. President Bush (2008) made similar claims about results from the main assessment: "Eighth graders set a record high for math scores last year."

14. As Fields (2008) points out, 9-year-olds gained 11 points between 1971 and 2004, or about .2 of a standard deviation. While some would consider this substantial, a .2 gain is traditionally labeled a "small" effect size. Seven of the eleven points came between 1999 and 2004, which could reflect a real improvement in reading or the impact of teaching to the test caused by increased state testing in recent years and NCLB-mandated testing at younger ages. In any event, the 2004 reading score for 9-year-olds is only 4 points above the level in 1980 (Perie, Moran, and Lutkus, 2005, p. 10). Thirteen-year-olds also gained much less and 17-year-olds did not gain at all. See data table for figure 1 in appendix.

15. To be sure, commentators recognized there was a problem at the high school level, but made references only in passing. NCTM (2005) was perhaps the most forthright, acknowledging directly that trends had been flat among high school students.

16. Gains were more striking on the *main* assessment. In the 1990s, fourth graders gained 13 points while eighth graders gained 10 points (Lee, Grigg, and Dion, op. cit., pp. 8, 24). The gains varied greatly by parental education, however, indicating the standards were not having a *general* impact (Campbell, Hombo, and Mazzeo, 2000, p. 117). The discrepant results of the main and long-term assessments appear unrelated to their content: panels of math educators and teachers judged their test items equally relevant with similar classroom coverage (Pellegrino, Jones, and Mitchell, 1999, p. 78). The gains on NAEP did not solely reflect real math improvement, but were partly caused by compositional changes in test takers and teaching to the test. Even in the 1990s, compositional effects were substantial. In the long-term trend assessment, the proportion of NAEP test takers with a college-graduate parent rose from about 40 percent to nearly 50 percent. If scores had remained constant, this changing mix alone would have produced about one-quarter of the gain (Campbell, Hombo, and Mazzeo, 2000, p. 117). One can show a similar impact in the main assessment using 1990–2000 eighth grade scores and parental education percentages (Braswell et al., 2001, p. 246).

17. As indicated in a paper written for The Brookings Institution (Stedman, 1998), the predictive validity of these levels is "undetermined—the connection between a given level of performance and future academic or economic success is likely more tenuous than claimed in the NAEP reports (Chelimsky, 1992, p. 4; Forsyth, 1991, pp. 3–9, 16)."

18. Reading and math data are in Grigg, Donahue, and Dion, 2007, pp. 5, 15. Science data are in Grigg, Lauko, and Brockway, 2006, p. 31. Writing data are in Salahu-Din, Persky, and Miller, 2008, p. 37. Civics data are in Lutkus and Weiss, 2007, p. 1. History data are in Lee and Weiss, 2007, p. 9. Economics data are in Mead and Sandene, 2007, p. 5. Geography data are in Weiss et al, 2002, p. 21.

19. Publicly released items and performance data are available on the web, http://nces.ed.gov/nationsreportcard/itmrls/startsearch.asp. Select age 17, mathematics, and 2004.

20. To measure national achievement thoroughly and accurately will require a return to NAEP's original practice of testing out-of-school teenagers.

21. The percentage reporting 5 or more hours a week had grown from 29 to 63 percent (NCES, 2007, p. 159).

22. This is the report of the National Research Council's Committee on the Evaluation of National and State Assessments of Educational Progress (Pellegrino, Jones, and Mitchell, 1999, p. 2). The study was supported by a contract to the National Academy of Sciences, so it is sometimes referred to as the NAS evaluation. A frontispiece notice explains that it was a project of the National Research Council (Pellegrino, Jones, and Mitchell, 1999, II).

23. The trend studies of the 1990s reported, for example, that the long-term reading assessment was based on that of 1984 and "most closely reflects the objectives developed for that assessment" (Campbell, Hombo, and Mazzeo, 1999, p. 85). The writing trend line began in 1984. Math and science were based on 1986 curricular and testing frameworks.

Bibliography

Advisory Panel on the Scholastic Aptitude Test Score Decline. (1977). *On further examination.* New York: College Entrance Examination Board.

Ahmann, J. (1979). National Achievement Profiles in Ten Learning Areas. *Educational Studies,* Winter, 351–364.

Aikin, W. (1942). *The story of the eight-year study.* New York: Harper.

American Federation of Teachers. (2005, July 14). Statement by Antonia Cortese, Executive Vice President, American Federation of Teachers, on the 2004 National Assessment of Educational Progress (NAEP) long-term trends in academic progress. Link reconfirmed August 23, 2008. http://www.aft.org/presscenter/releases/2005/071405.htm

Anderson, L., Jenkins, L., Leming, J., MacDonald, W., Mullis, I., Turner, M., and Wooster, J. (1990). The civics report card: *Trends in achievement from 1976 to 1988 at ages 13 and 17; achievement in 1988 at grades 4, 8, and 12.* Princeton, NJ: Educational Testing Service.

Applebee, A., Langer, J., and Mullis, I. (1987). *Literature & U.S. history: The instructional experience and factual knowledge of high school juniors.* Princeton, NJ: Educational Testing Service.

Applebee, A., Langer, J., and Mullis, I. (1988). *Literature & U.S. history: The instructional experience and factual knowledge of high school juniors.* Princeton, NJ: Educational Testing Service.

Applebee, A., Langer, J., and Mullis, I. (1989). *Crossroads in American Education.* Princeton, NJ: Educational Testing Service.

Applebee, A., Langer, J., Mullis, I., and Jenkins, L. (1990). *The writing report card, 1984–1988.* Princeton, NJ: NAEP.

Augustus F. Hawkins-Robert T. Stafford Elementary and Secondary School Improvement Amendments of 1988. Public Law 100–297. ERIC Document Reproduction No. ED 307 960.

Ballator, N., Farnum, M., and Kaplan, B. (1999). *NAEP 1996 trends in writing: fluency and writing conventions.* NCES 1999–456. Washington, DC: U.S. Government Printing Office.

Bennett, S. and Kalish, N. (2006). *The case against homework: How homework is hurting children and what parents can do about it.* New York: Three Rivers Press.

Berliner, D. and Biddle, B. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools.* New York: Addison-Wesley.

Berliner, D. and Biddle, B. (1996). Making molehills out of molehills: Reply to Lawrence Stedman's review of the manufactured crisis. *Education Policy Analysis Archives* 4(3). Link reconfirmed August 23, 2008. http://epaa.asu.edu/epaa/v4n3.html

Bracey, G. (1992). The second Bracey report on the condition of public education. *Phi Delta Kappan* 74(2): 104–117.

Bracey, G. (1995, December 22). U.S. students: Better than ever. *Washington Post,* p. A19.

Braswell, J., Lutkus, A., Grigg, W., Santapau, S., Tay-Lim, B., and Johnson, M. (2001). *The Nation's Report Card: Mathematics 2000.* Washington, DC: Government Printing Office.

Bush, G. (2008). President Bush discusses the No Child Left Behind Act. Press release. Link reconfirmed August 23, 2008. http://www.whitehouse.gov/news/releases/2008/01/20080107-2.html

Cahalan, M., Ingels, S., Burns, L., Planty, M., and Daniel, B. (2006). *United States high school sophomores: A twenty-two year comparison, 1980–2002.* NCES 2006–327. Washington, DC: U.S. Government Printing Office.

Campbell, J., Hombo, C., and Mazzeo, J. (2000). *NAEP 1999 Trends in academic progress: Three decades of student performance.* NCES 2000–469. Washington, DC: Government Printing Office.

Campbell, J., Reese, C., O'Sullivan, C., and Dossey, J. (1996). *NAEP 1994 trends in academic progress.* Washington, DC: U.S. Department of Education.

Campbell, J., Voelkl, K., and Donahue, P. (1998). *NAEP 1996 trends in academic progress. Addendum. Achievement of U.S. students in science, 1969 to 1996; mathematics, 1973 to 1996; reading, 1971 to 1996; writing, 1984 to 1996. Revised.* Washington, DC: U.S. Government Printing Office. ERIC Document Reproduction No. ED 424 269.

Carpenter, T., Lindquist, M., Brown, C., Kouba, V., Silver, E., and Swafford, J. (1988). Results of the fourth NAEP assessment of mathematics. *Arithmetic Teacher* (December): 38–41.

Carroll, J. (1988). The NAEP reading proficiency scale is not a fiction: A reply to McLean and Goldstein. *Phi Delta Kappan,* June, pp. 761–764.

Cavanagh, S. (2005, March 16). Board studies release of individual NAEP results. *Education Week* 24(7): 27–28.

Cavanagh, S. (2008, March 11). 11 states poised to pilot national test for seniors. *Education Week* 6.

Center for Public Education. (2008). *The proficiency debate: A guide to NAEP achievement levels.* Link reconfirmed August 22, 2008. http://www.centerforpubliceducation.org/site/c.kjJXJ5MPIwE/b.4175355/k.9E78/The_proficiency_debate_A_guide_to_NAEP_achievement_levels.htm

Chall, J. (1995). *Learning to Read: The great debate.* New York: Harcourt Brace.

Chelimsky, E. (1992). *National Assessment Governing Board (NAGB) achievement levels. Interim letter report.* Washington, DC: General Accounting Office.

Cizek, G. (1993). *Reactions to National Academy of Education report "Setting performance standards for students achievement.* Washington, DC: National Assessment Governing Board. ERIC Document Reproduction No. ED 360 397.

Copperman, P. (1978). *The literacy hoax: The decline of reading, writing, and learning in the public schools and what we can do about it.* New York: William Morrow.

Cremin, L. (1961). *The transformation of the school.* New York: Vintage Books.

Dossey, J., Mullis, I., Lindquist, M., and Chambers, D. (1988). *The mathematics report card: Are we measuring up?* Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1985). *The reading report card: Progress toward excellence in our schools. Trends in reading over four national assessments, 1971–1984.* Report No. 15–R–01. Princeton, NJ: Educational Testing Service. ERIC Document Reproduction No. ED 264 550.

Fields, R. (2008). Personal communication on original manuscript.

Finder, M. (2004). *Educating America: How Ralph W. Tyler taught America to teach.* London: Praeger.

Fitzharris, L. (1993). *An Historical Review of the National Assessment of Educational Progress from 1963 to 1991.* Dissertation. University of South Carolina.

Flesch, R. (1981). *Why Johnny still can't read: A new look at the scandal of our schools.* New York: Harper & Row.

Forsyth, R. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice* 10(3): 3–9.

Fuller, B., Wright, J., Gesicki, K., and Kang, E. (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher* 36(5): 268–278.

Glass, G. (2003, February). Standards and Criteria Redux. Link reconfirmed August 22, 2008. http://glass.ed.asu.edu/gene/papers/standards/ Updated version of Glass, G. (1978). Standards and criteria. *Journal of Educational Measurement* 15: 237–261.

Goodman, K. (Ed.). (2004). *In defense of good teaching: What teachers need to know about the "reading wars."* York, ME: Stenhouse Publishers.

Greenwald, E., Persky, H., Campbell, J., and Mazzeo, J. (1999). *NAEP 1998 writing report card for the nation and the states.* Washington, DC: U.S. Department of Education.

Grigg, W., Donahue, P., and Dion, G. (2007). *The nation's report card: 12th grade reading and mathematics 2005.* NCES 2007–468. Washington, DC: U.S. Government Printing Office.

Grigg, W., Lauko, M., and Brockway, D. (2006). *The nation's report card: Science 2005.* NCES 2006–466. Washington, D.C.: U.S. Government Printing Office.

Hammack, D., Hartoorian, M., Howe, J., Jenkins, L., Levstik, L., MacDonald, W., Mullis, I., and Own, E. (1990). *The U.S. history report card: The achievement of fourth-, eighth- and twelfth-grade students in 1988 and trends from 1986 to 1988 in the factual knowledge of high-school juniors.* Princeton, NJ: Educational Testing Service.

Hand, H. (1965). National assessment viewed as the camel's nose. *Phi Delta Kappan* 47: September 8–13.

Hirsch, E.D. (1987). *Cultural literacy.* Boston: Houghton Mifflin Company.

Ho, A. and Haertel, E. (n.d.). Apples to apples? The underlying assumptions of state-NAEP comparisons. Link reconfirmed August 22, 2008. http://www.ccsso.org/content/PDFs/Ho Haertel CCSSO Brief2 Final.pdf

Hoff, David J. (2007, June 20). State tests show gains since NCLB; Report cautions against crediting education law" *Education Week* 26(39): 1, 20.

Institute for Language and Education Policy. (2007). Would you buy a used law from this woman? Link reconfirmed August 23, 2008. http://www.elladvocates.org/nclb/spellings2.html

Itzkoff, S. (1994). *The decline of intelligence in America: A strategy for national renewal.* Westport, CT: Praeger.

Jones, L. (1996). A history of the National Assessment of Educational Progress and some questions about its future. *Educational Researcher* October: 15–22.

Kaestle, C., Damon-Moore, H., Stedman, L., Tinsley, K., and Trollinger, W. (1991). *Literacy in the United States: Readers and reading since 1880.* New Haven: Yale University Press.

Kliebard, H. (1995). *The Struggle for the American Curriculum: 1893–1958.* London: Routledge.

Kohn, A. (2006). *The Homework Myth: Why our kids get too much of a bad thing.* Philadelphia: Da Capo Books.

Krashen, S. (2006, September 15). Did NCLB raise fourth grade NAEP scores? A response to Shanahan & Hynd Shanahan. Links reconfirmed August 23, 2008. http://sdkrashen.com/pipermail/krashen_sdkrashen.com/2006-September/000603.html and http://www.tcrecord.org/Discussion.asp?i=3&vdpid=2623&aid=2&rid=12696&dtid=0

Krug, E. (1972). *The shaping of the American high school, volume 2, 1920–1941.* Madison, WI: The University of Wisconsin Press.

Langer, J., Campbell, J., Neuman, S., Mullis, I., Persky, H., and Donahue, P. (1995). *Reading assessment redesigned: Authentic texts and innovative instruments in NAEP's 1992 survey.* Report No. 23–FR–07. Washington, DC: U.S. Government Printing Office.

Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends.* Cambridge, MA: The Civil Rights Project at Harvard University.

Lee, J. and Weiss, A. (2007). *The nation's report card: U.S. history 2006.* NCES 2007–474. Washington, DC: U.S. Government Printing Office.

Lee, J., Grigg, W., and Dion, G. (2007). *The nation's report card: Mathematics 2007.* NCES 2007–494. Washington, DC: U.S. Government Printing Office.

Lee, J., Grigg, W., and Donahue, P. (2007). *The nation's report card: Reading 2007.* NCES 2007–496. Washington, DC: U.S. Government Printing Office.

Lutkus, A. and Weiss, A. (2007). *The nation's report card: Civics 2006.* NCES 2007–476. Washington, DC: U.S. Government Printing Office.

McLean, L. and Goldstein, H. (1988). The U.S. national assessments in reading: Reading too much into the findings. *Phi Delta Kappan* January: 369–372.

Mead, N. and Sandene, B. (2007). *The Nation's Report Card: Economics 2006.* NCES 2007–475. Washington, DC: U.S. Government Printing Office.

Mullis, I. and Jenkins, L. (1990). *The reading report card, 1971–1988.* Princeton, NJ: Educational Testing Service.

Mullis, I. and Jenkins, L. (1988). *The science report card: Elements of risk and recovery. Trends and achievement based on the 1986 national assessment*. Princeton, NJ: Educational Testing Service.

Mullis, I., Dossey, J., Foertsch, M., Jones, L., and Gentile, C. (1991). *Trends in academic progress: Achievement of U.S. students in science, 1969–70 to 1990; mathematics, 1973 to 1990; reading, 1971 to 1990; and writing, 1984 to 1990.* Washington, DC: U.S. Government Printing Office. ERIC Document Reproduction No. ED 338 720.

Mullis, I., Owen, E., and Phillips, G. (1990.) *America's challenge: Accelerating academic achievement.* Report No. OV-01. Princeton, NJ: Educational Testing Service.

National Assessment Governing Board. (1996, August 2). *Policy statement on redesigning the National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board.

National Assessment Governing Board. (2002, May 18). *National Assessment Governing Board long-term trend policy statement.* Washington, DC: National Assessment Governing Board.

National Assessment of Educational Progress. (1970). *Report 2—Citizenship: National results—partial. Observations and commentary of a panel of reviewers.* Washington, DC: U.S. Government Printing Office. ERIC Document Reproduction No. ED 049 112.

National Assessment of Educational Progress. (1976). *Reading in America: A Perspective on two assessments.* NAEP–06–R–01. Washington, DC: U.S. Government Printing Office. ERIC Document Reproduction No. ED 128 785.

National Assessment of Educational Progress. (1978). *Changes in political knowledge and attitudes, 1969–76.* Citizenship/social studies report No. 07–CS–02. Denver: Education Commission of the States.

National Assessment of Educational Progress. (1981). *Three national assessments of reading: Changes in performance, 1970–1980.* Report 11–R–01. Denver: Education Commission of the States.

National Assessment of Educational Progress. (1982). *The reading comprehension of American youth.* Report Number 11–R–02. Denver: Education Commission of the States.

National Assessment of Educational Progress. (1983). *The third national mathematics assessment: Results, trends, and issues.* Report No. 13–MA–01. Denver: Education Commission of the States.

National Assessment of Educational Progress. (1983). *The third national mathematics assessment: Results, trends, and issues.* Report No. 13–MA–01. Denver: Education Commission of the States.

National Assessment of Educational Progress. (2008). Publicly released sample items available on the web. Select Long-term Trend, subject, 2004, age 17. Link reconfirmed August 23, 2008. http://nces.ed.gov/nationsreportcard/itmrls/startsearch.asp

National Assessment of Educational Progress. (2009). The history of NAEP contractors. Link confirmed January 13, 2009. http://nces.ed.gov/nationsreportcard/contracts/history.asp

National Center for Education Statistics. (1993). *Digest of education statistics 1993.* U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

National Center for Education Statistics. (2003). *Digest of education statistics 2003.* U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

National Center for Education Statistics. (2005). Long-term trend major results. Link reconfirmed August 23, 2008. http://nces.ed.gov/nationsreportcard/ltt/results2004/

National Center for Education Statistics. (2006). *Digest of education statistics 2006.* U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

National Center for Education Statistics. (2007). *The condition of education 2007.* NCES 2007–064. Washington, DC: U.S. Government Printing Office.

National Center for Education Statistics. (2008a). *Chronology of National Assessment of Educational Progress (NAEP) assessments from 1969 to 2006.* Link reconfirmed August 22, 2008. http://nces.ed.gov/nationsreportcard/about/assesshistory.asp

National Center for Education Statistics. (2008b). NAEP overview. *NAEP assessments: Main and long-term trend.* Link reconfirmed August 23, 2008. http://nces.ed.gov/nationsreportcard/about/#mainlongterm

National Center for Education Statistics. (2008c). *Schedule for the state and National Assessment of Educational Progress (NAEP) from 2007 to 2017.* Link reconfirmed August 22, 2008. http://nces.ed.gov/nationsreportcard/about/assessmentsched.asp

National Center for Education Statistics. (2008d). *What are the differences between long-term trend NAEP and main NAEP?* Link reconfirmed August 23, 2008. http://nces.ed.gov/nationsreportcard/about/ltt_main_diff.asp

National Commission on Excellence in Education. (1983). *A Nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

National Council of Teachers of English. (1996). *Standards for the English language arts.* Urbana, IL: NCTE.

National Council of Teachers of Mathematics. (2004, January/February). Rise in NAEP math scores coincides with NCTM standards. *News Bulletin*. Available at http://www.nctm.org/news/release.aspx?id=766.

National Council of Teachers of Mathematics. (2005, July 14). Trend NAEP reports math scores up for 9- and 13-year-olds; 17-year-olds steady. Press release. Link reconfirmed August 23, 2008. http://www.nctm.org/news/content.aspx?id=714

National Education Goals Panel. (1995). *The national education goals report: Building a nation of learners 1995.* Washington, DC: U.S. Government Printing Office.

National Education Goals Panel. (1999a). *Reading achievement state by state, 1999.* Washington, DC: U.S. Government Printing Office.

National Education Goals Panel. (1999b). *The national education goals report: Building a nation of learners 1999.* Washington, DC: U.S. Government Printing Office.

National Endowment for the Arts. (2004). *Reading at risk: A survey of literary reading in America.* Research Report #46. Washington, DC: NEA.

National Endowment for the Arts. (2007). *To Read or Not To Read: A Question of National Consequence.* Research Report #47. Washington, DC: NEA.

National Endowment for the Arts. (2009). *Reading on the Rise: A New Chapter in American Literacy.* Washington, DC: NEA.

No Child Left Behind Act of 2001. Public Law 107–110. 115 Stat. 1425. Link reconfirmed August 23, 2008. http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf

O'Sullivan, C., Lauko, M., Grigg, W., Qian, J. and Zhang, J. (2003). *The nation's report card: Science 2000.* NCES 2003–453. Washington, DC: U.S. Department of Education.

Orfield, G. (2006). *Forward to J. Lee, Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends.* Cambridge, MA: The Civil Rights Project at Harvard University.

Pellegrino, J., Jones, L., and Mitchell, K. (Eds.) (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress.* Report of the Committee on the Evaluation of National and State Assessments of Educational Progress, Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy Press.

Perie, M., Grigg, W., and Dion, G. (2005). *The nation's report card: Mathematics 2005.* NCES 2006–453. Washington, DC: U.S. Government Printing Office.

Perie, M., Moran, R., and Lutkus, A. (2005). *NAEP 2004 Trends in academic progress: Three decades of student performance in reading and mathematics.* NCES 2005–464. Washington, DC: Government Printing Office.

Persky, H., Reese, C., O'Sullivan, C., Lazer, S., Moore, J., and Shakrani, S. (1996). *NAEP 1994 geography report card.* Washington, DC: U.S. Department of Education.

Ravitch, D. (1995). *National standards in American education: A citizen's guide.* Washington, DC: The Brookings Institution.

Reese, C., Miller, K., Mazzeo, J., and Dossey, J. (1997). *NAEP 1996 mathematics report card for the nation and the states.* Washington, DC: U.S. Department of Education.

Rich, M. (2007, November 19). Study Links Drop in Test Scores to a Decline in Time Spent Reading. New York Times, Arts Section. Link reconfirmed August 23, 2008. http://www.nytimes.com/2007/11/19/arts/19nea.html

Rich, M. (2008, July 27). Literacy Debate: Online, R U Really Reading? *New York Times.* Book Section. Link reconfirmed August 23, 2008. http://www.nytimes.com/2008/07/27/books/27reading.html?_r=2&hp&oref=slogin&oref=slogin

Sadker, M. and Sadker, D. (1994). *Failing at fairness: How our schools cheat girls.* New York: Simon & Schuster.

Salahu-Din, D., Persky, H., and Miller, J. (2008). *The nation's report card: Writing 2007.* NCES 2008–468. Washington, DC: U.S. Government Printing Office.

Shanahan, T. and Hynd-Shanahan, C. (2006, September 5). A good start is not enough: what it will take to improve adolescent literacy. Posted on Teachers College Record tcrecord.org. Link reconfirmed August 23, 2008. http://www.tcrecord.org/Content.asp?ContentID=12696

Shanker, A. (1991). Do private schools outperform public schools? *American Educator* 15(2): 8–15+.

Smith, M. (2007). *Leaving NCLB renewal behind.* Link reconfirmed August 23, 2008. http://www.educationevolving.org/pdf/mikesmithoped.pdf

Sommers, C. (1994). *Who stole feminism? How women have betrayed women.* New York: Simon & Schuster.

Stake, R. (2007). NAEP, Report Cards and Education: A Review Essay. *Education Review* 10(1): 1–22.

Stedman, L. (1995). The new mythology about the status of U.S. schools. *Educational Leadership* 52(5): 80–85.

Stedman, L. (1996a). Respecting the evidence: The achievement crisis remains real. [Review of *The manufactured crisis*.] *Education Policy Analysis Archives* 4(7). Link reconfirmed August 23, 2008. http://epaa.asu.edu/epaa/v4n7.html

Stedman, L. (1996b). The achievement crisis is real. [Review of *The manufactured crisis*.] *Education Policy Analysis Archives* 4(1). Link reconfirmed August 23, 2008. http://epaa.asu.edu/epaa/v4n1.html

Stedman, L. (1997). International achievement differences: An assessment of a new perspective. *Educational Researcher* 26(3): 4–15.

Stedman, L. (1998). An assessment of the contemporary debate over U.S. achievement. In D. Ravitch (Ed.), *Brookings papers on education policy: 1998* (pp. 53–121). Washington, DC: The Brookings Institution.

Stedman, L. (2003). U.S. educational achievement in the 20th century: Brilliant success and persistent failure. In R. Weissberg, H. Walberg, M. O'Brien, and C. Bartels-Kuster (Eds.), *Long-term trends in the well-being of children and youth* (Chapter 3). Washington, DC: Child Welfare League of America Press.

Stedman, L. and Kaestle, C. (1986). *An investigation of crude literacy, reading performance, and functional literacy in the United States, 1880 to 1980.* Program Report 86–23. Madison, WI: Wisconsin Center for Education Research at the University of Wisconsin.

Stedman, L. and Kaestle, C. (1987). Literacy and reading performance in the United States, from 1880 to the present. *Reading Research Quarterly* XXII(1): 8–46.

Stedman, L. and Kaestle, C. (1985). The test score decline is over: Now what? *Phi Delta Kappan* 67(3): 204–210.

Stedman, L. and Smith, M. (1983). Recent reform proposals for American education. *Contemporary Education Review* 2(2): 85–104.

Stedman, L. and Smith, M. (1983). Recent reform proposals for American education. *Contemporary Education Review* 2(2): 85–104.

Steinberg, L., Brown, B., and Dornbusch, S. (1996). *Beyond the classroom: Why school reform has failed and what parents can do.* New York: Simon & Schuster.

Sykes, C. (1995). *Dumbing down our kids.* New York: St. Martin's Press.

Taylor, D. (1998). *Beginning to read and the spin doctors of science: The political campaign to change America's mind about how children learn to read.* Urbana, IL: National Council of Teachers of English.

Thimmesch, N. (Ed.) (1984). *Aliteracy: People who can read but won't.* London: American Enterprise Institute for Public Policy Research.

Thompson, B. (2009, January 12). Unexpected Twist: Fiction Reading Is Up Survey Shows Reversal Of Longstanding Trend. *Washington Post,* p. C01.

Tukey, J., Abelson, R., Coffman, W., Jones, L., and Mosteller, F. (1971). *National assessment report 41969–1970 science: Group results for sex, region, and size of community.* Washington, DC: Government Printing Office.

U.S. Department of Education. (2005, July 14). Spellings hails new national report card results: Today's news "proof that No Child Left Behind is working." Press release. Link reconfirmed August 23, 2008. http://www.ed.gov/news/pressreleases/2005/07/07142005.html

Vinovskis, M. (1998). *Overseeing the nation's report card: The creation and evolution of the National Assessment Governing Board.* Washington, DC: U.S. Department of Education.

Weiss, A., Lutkus, A., Grigg, W., and Niemi, R. (2001). *The next generation of citizens: NAEP civics assessments–1988 and 1998.* NCES 2001–452. Washington, DC: U.S. Government Printing Office.

Weiss, A., Lutkus, A., Hildebrant, B., and Johnson, M. (2002). *The nation's report card: Geography 2001.* NCES 2002–484. Washington, DC: U.S. Department of Education.