

White Paper on 50 Years of NAEP Use: Where NAEP Has Been and Where It Should Go Next

Ina V.S. Mullis, PhD
Boston College

June 2019
Commissioned by the NAEP Validity Studies (NVS) Panel

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics. Points of view or opinions expressed in this paper do not necessarily represent the official positions of the U.S. Department of Education or the American Institutes for Research.

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Peter Behuniak
Criterion Consulting, LLC

Scott Norton
Council of Chief State School Officers

Jack Buckley
American Institutes for Research

James Pellegrino
University of Illinois at Chicago

James R. Chromy
Research Triangle Institute (retired)

Gary Phillips
American Institutes for Research

Phil Daro
*Strategic Education Research Partnership (SERP)
Institute*

Lorrie Shepard
University of Colorado Boulder

Richard P. Durán
University of California, Santa Barbara

David Thissen
University of North Carolina, Chapel Hill

David Grissmer
University of Virginia

Gerald Tindal
University of Oregon

Larry Hedges
Northwestern University

Sheila Valencia
University of Washington

Gerunda Hughes
Howard University

Denny Way
College Board

Ina V.S. Mullis
Boston College

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Grady Wilburn
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS) Panel
American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
E-Mail: fstancavage@air.org

CONTENTS

EVOLUTION OF THE PRIMARY USES OF NAEP	1
Origins of the Idea for NAEP	1
NAEP 1963 to 1969: The Gestation Period	2
NAEP 1970 to 1983: The Ebb and Flow of Early NAEP	3
1983 to 1988: ETS's New Design for a New Era	5
1984 to 2002: The Trial State Assessments and the Creation of the National Assessment Governing Board	6
2002: The Current NAEP Legislation	9
2002 to 2018: Assessments Conducted	11
2002 to 2018: NAEP Report Cards Provided by NCES	12
2009 to 2018: NAEP Transitions to Digital-Based Assessments	13
ADDITIONAL USES OF NAEP	16
Using NAEP to Improve State Proficiency Standards	16
NAEP Special Studies	17
NAEP Data Explorer and Availability of Data for Secondary Analysis	18
Availability of Frameworks and Questions	18
National Assessment Governing Board Studies and Reports	19
Using NAEP Data to Improve NAEP	19
NAEP's Potential to Inform Policy	20
Next Steps for NAEP	22
REFERENCES	25

EVOLUTION OF THE PRIMARY USES OF NAEP

Now that the National Assessment of Educational Progress (NAEP) has reached its 50th birthday, it seems appropriate to consider its history and celebrate its many accomplishments. In particular, this paper reflects on factors influencing NAEP's use as the Nation's Report Card, including legislative- and funding-driven changes to both the data collected (subject areas and populations assessed) and the manner in which the data are reported. From 1969 through 2018, there also have been different approaches to dissemination efforts, access to the data for further research, and the degree of support given to potential users. Most recently, there is the transition to digital-based assessments in 2017 and the promise that holds for keeping NAEP innovative and making its results even more useful.^{1 2}

Origins of the Idea for NAEP

Historically, the fundamental purpose and use of NAEP was specified about 100 years before its realization in 1969. The idea of a Nation's Report Card dates back to 1867, as explained in the *NAEP Primer* (Beaton et al., 2011):

In 1962, President Kennedy appointed Francis Keppel as head of the Office of Education, which was then part of the Department of Health, Education, and Welfare. As Keppel told the story, he went to Washington and wondered what the duties of the Commissioner of Education were, so he looked up the 1867 law that authorized the Office. He found that the Office was to **report annually on the progress of students in the United States** [emphasis added]. He marveled at the fact that, in nearly a century, the Office had never done so.

In March of 1963, Keppel asked Ralph Tyler, then Director of the Institute for Advanced Studies in the Behavioral Sciences, for his suggestions for **measuring school quality** [emphasis added]. Tyler replied shortly, and thus the germ of the idea that was to grow into the National Assessment of Educational Progress was sown. In its early days, Professor John Tukey of Princeton was the technical leader of the project. As the project grew, it came under the administrative leadership of the Education Commission of the States (ECS).

ECS was the operating arm for the Compact for Education, a state organization proposed by James Bryant Conant at about the same time as the idea of NAEP (Education Commission of the States, n.d.). Established in 1967, the Compact for Education was endorsed by all 50 states, approved by Congress, and aimed to promote states working together to focus on national education issues. As described in the next section, ECS was given responsibility for managing NAEP in 1969.

¹ The author worked on NAEP for parts of this time period: at the Education Commission of the States from 1974 to 1983 and the Educational Testing Service from 1984 to 1994. She has been a member of the National Center for Education Statistics (NCES) NAEP Validity Studies (NVS) Panel since 1995.

² This paper has benefited enormously from constructive comments and corrections by a number of reviewers. Their assistance is gratefully acknowledged. Any remaining errors are the responsibility of the author. Views and opinions expressed are the author's own and do not necessarily represent those of the U.S. Department of Education, NCES, or the National Assessment Governing Board.

NAEP 1963 to 1969: The Gestation Period

A 1974 article in *PTA Magazine* stated that NAEP had the longest gestation period of any education research project on record—6 years, from 1963 to 1969 (Hazlett, 1974). In retrospect, this was not long considering the enduring importance of the work that was done.

To begin, Tyler responded to Keppel's call for a measure of school quality in a memorandum that recommended a periodic index that could be used to monitor educational improvement and inequity of educational levels among regions or states; Keppel then obtained funds from the Carnegie Corporation, and planning proceeded (Hazlett, 1974). From 1963 to 1969, the founders of NAEP developed an innovative, new, large-scale measurement technology; specified the assessments to be conducted; and negotiated an arrangement for managing the project.

More specifically, from 1963 through 1964, under Tyler's leadership, there were two initial planning meetings, followed by funding for the Exploratory Committee for the Assessment of Progress in Education (ECAPE). In 1965, ECAPE created a Technical Advisory Committee (TAC) chaired by John Tukey, became incorporated as a nonprofit corporation, was awarded nearly \$3 million (\$2.3 million from the Carnegie Corporation and nearly \$500,000 from the Ford Foundation), and began to review proposals from contractors. In 1968, ECAPE reformed as the Committee for the Assessment of Progress in Education (CAPE), and, in April 1969, CAPE conducted the first NAEP assessment of in-school 17-year-olds in citizenship, science, and writing. This was quickly followed, in December 1969, by an assessment of 13- and 9-year-olds in the same subject areas. Concurrently, TAC (renamed the Analysis Advisory Committee [ANAC]) wrote reports of the science results as models, so, by 1970, NAEP was under way (Jones, 2004).

Once CAPE had implemented the NAEP design, an important part of the task of getting it established was deciding about future oversight. The United States Office of Education (USOE) had begun providing substantial support to NAEP in 1969 (\$1 million), but the country was still reverberating from Sputnik (as well as being in the midst of the Vietnam War and Civil Rights protests), so there were concerns about the role of the federal government in general, and specific concerns about NAEP being a first step in dismantling traditional state and local control of education.

As explained by then Secretary of Health, Education, and Welfare John Gardner in an interview with Lyle V. Jones and Ingram Olkin (2004), “the very thing that almost shot us down” was the plan “to create a mechanism to evaluate the quality of the school, rather than to focus the assessment on the individual student” (pp. 114–115). The NAEP founders had naively thought schools would welcome the idea of an “evaluation of how the nation was doing, how states were doing, and how school districts were doing,” but the American Association of School Administrators “didn't want anybody to measure them and they fought hard” (p. 116). Also, the National Education Association (NEA) decided to back the school administrators. In 1967, NEA passed a resolution to withhold cooperation from NAEP. With the education establishment lining up against NAEP, compromises were made—that “NAEP would not be used for any direct comparison of schools, that it would be used for assessing year-to-year progress, not district versus district” (p. 117).

In 1969, following a public meeting for comments, ECS, as noted above, assumed management of NAEP and created the first national Assessment Policy Committee (APC) to

provide advice to the project. With a state organization at the helm, it was felt that fears of national testing and a national curriculum could be laid to rest (Hazlett, 1974).

NAEP 1970 to 1983: The Ebb and Flow of Early NAEP

In the early 1970s, NAEP was fielding state-of-the-art assessments in an array of subject areas. From the beginning, Tyler stressed that what had been learned was as important as what had not, and that current standardized tests could not be used, primarily because they were designed to differentiate among students rather than describe what students could do (Hazlett, 1974). Tyler's idea was that performance data on the individual items had merit, and, consequently, the results were reported in terms of the percentages of students answering each item correctly (*p*-values). Progress was monitored by changes in these percentages over time. The reports also were framed in accordance with ECS's policy to not interpret or speculate on the results; thus, the data were considered "census-like," and no presumed causative factors were included in the analyses or reporting (Hazlett, 1974).

Early NAEP had several other important characteristics, some of which are still recognizable today:

- It was based on sampling students rather than a census approach. This was one reason it was feasible to conduct a national assessment.
- It covered 10 subject areas, each scheduled to be assessed on a 5-year cycle (twice in 10 years). The designers felt the content coverage should be as broad as possible, measuring all major educational goals (that is, all topics that, on average, accounted for at least 15–20% of instructional time in U.S. schools). By summarizing the results of all 10 areas on a 5-year cycle, NAEP could produce a document that described the overall educational health of the country and the strengths and weaknesses across subject areas (Hazlett, 1974).
- The learning objectives to be assessed for each subject area were developed by broadly representative groups of scholars, educational practitioners, and thoughtful lay citizens.
- The populations assessed included 9-, 13-, and 17-year-olds attending school as well as 17-year-olds not attending school and young adults aged 26 to 35. The young adult group was intended to provide representation of the educational accomplishments of students who had finished their schooling.
- Innovative assessment approaches were adopted, including individual interviews and group problem solving as well as performance assessments (e.g., music, art, occupational development). The planners emphasized that assessment procedures should be varied beyond paper-and-pencil tests.

In her chapter of *The Nation's Report Card: Evolution and Perspectives*, Dorothy Gilford (2004), then NCES director, describes NAEP from 1971 to 1974. In 1971, USOE transferred NAEP to NCES, and, by 1972–73, funding from USOE to ECS had risen to \$6 million. According to Gilford, ECS had, by that time, hired a special staff dedicated to NAEP, with ANAC providing technical direction, and the NAEP program was functioning quite effectively. Between the fall of 1971 and 1974, assessments were conducted in five subjects (including two assessments in science), and a large number of reports were written.

Gilford also reported that the NAEP program engaged in a number of activities to promote and encourage use of NAEP methods and data. There was considerable interest in the new large-scale assessment technology pioneered by NAEP, and NCES—working through ECS—provided services to state and local education systems in three ways: (1) adapting the NAEP model to local needs, (2) providing technical assistance, and (3) holding workshops and seminars to exchange ideas. Several states developed their own assessments. Also, because external data use was envisioned to focus on using item-by-item results to improve curriculum and instruction, NCES also encouraged ECS to award small contracts to teacher organizations (e.g., National Science Teachers Association, National Council for the Social Studies, and National Council of Teachers of Mathematics) so that they could review the NAEP results for their memberships. According to Gilford, prior to the changes introduced by USOE in 1974 (see below), NAEP was quite effective in meeting its overarching goal “to provide information useful to educational decision-makers and practitioners in identifying problems, setting priorities, and determining progress” (p. 176).

Unfortunately, despite this track record of effectiveness, NAEP’s budget was halved to \$3 million in 1974. ECS was awarded the next contract, but many activities were curtailed or discontinued, such as assessing out-of-school populations and fielding innovative approaches to assessment. Not surprisingly, after a few years of the reduced funding and fewer assessment activities, the U.S. General Accounting Office (GAO) reported in 1976 that NAEP was not perceived as being particularly useful to policymakers and educators (U.S. General Accounting Office, 1976).

In 1978, sponsorship of NAEP and control over NAEP’s major activities was transferred within USOE to the newly established National Institute of Education (NIE). ECS was awarded a 5-year contract, but had less latitude and a smaller budget for conducting NAEP. The NIE legislation specified the membership of APC in terms of the constituencies that were to be represented, and it formalized the serious reduction in NAEP’s assessment responsibilities. In particular, as described below, assessments every year were reduced to “once every five years,” and the number of subjects to be assessed was reduced from 10 subjects to three.

NAEP was to:

- Collect and report, at least once every 5 years, data on the performance of students at various age or grade levels in reading, writing, and mathematics.
- Report on changes in the knowledge and skills of such students over a period of time.
- Conduct special assessments of other educational areas as the need for additional national information arose.
- Provide technical assistance to states and local agencies interested in adopting national assessment objectives, primarily pertaining to the basic skills of reading, mathematics, and communication.

In 1980, President Reagan was elected on a platform centered on reducing government, including the idea of abolishing the U.S. Department of Education (ED). Amid domestic budget reductions and concerns that NAEP’s data were not being widely used, Frederic Mosher of the Carnegie Foundation, together with his counterparts at both the Spencer and Ford Foundations, funded an evaluation of whether there was still a need for NAEP. The

study, completed in 1982 by Willard Wirtz and Archie Lapointe, found NAEP had little use and offered 13 recommendations for a more effective national assessment.

The next 5-year NAEP contract was scheduled for renewal in 1983. ECS had managed NAEP since 1969 and had no competition through the years. There was, however, competition for the 1983 award, including a bid from the Educational Testing Service (ETS), where Gregory Anrig, a member of the Wirtz and Lapointe (1982) study group, was president.

1983 to 1988: ETS's New Design for a New Era

In 1983, the Secretary of Education's National Commission on Excellence in Education released *A Nation at Risk: The Imperative for Educational Reform*, an influential and widely circulated report about the serious failings of American education. As described in the report, the Commission relied on five major sources of information, none of which were NAEP, although several NAEP findings were included (National Commission on Excellence, 1983). Nevertheless, 1983 also saw the beginning of major changes in NAEP, when ETS was awarded the \$3.88 million contract for NAEP's design, analysis, and reporting, with Westat, Inc., as the subcontractor for sampling and data collection.

ETS won the NAEP competition on the basis of bringing new ideas to NAEP and making the results more useful to policymakers. In a *New Design for a New Era* by Messick, Beaton, and Lord (1983), ETS described innovative methods and procedures for NAEP, such as balanced incomplete block designs for matrix sampling, item response theory (IRT) scaling, and extensive use of student- and school-level background data to frame the achievement results for policy analysis. ETS made Archie Lapointe the NAEP executive director and appointed new members to APC and TAC, still, however, including John Tukey on the latter.

During its first 5 years, ETS conducted assessments in 1984, 1986, and 1988 that covered each of the legislated subject areas at least once as well as several other subjects. ETS obtained additional funding to broaden the range of assessments, including the young adult literacy study of 16- to 25-year-olds; computer competence at Grades 3, 7, and 11; and geography, which was funded by the National Geographic Society.

The objectives and items for the assessments were developed at ETS under the guidance of committees of subject area experts. The results were analyzed using the new IRT scaling technology, which provided the basis for a new reporting approach. Results for the nation and population subgroups, as well as trends over time, were summarized on numerical achievement scales (from 0 to 500). The transition to IRT scaling methods from Tyler's reliance on percentage correct on individual items was fundamental in enabling trend measurement from assessment to assessment.

Also, during this time, several efforts were undertaken to increase NAEP's visibility and utility. At the suggestion of APC member Governor Lamar Alexander, NAEP was rebranded as the "The Nation's Report Card" to signal its relevance to a wider variety of audiences. Trend reporting gave NAEP greater traction with the media, and the interpretability of scores was improved by "anchoring" the scales in each subject area with descriptions of what students knew and could do at various points on the scale.

As Executive Director Archie Lapointe described in his chapter in *The Nation's Report Card: Evolution and Perspectives* (Lapointe, 2004), there also was a concerted campaign to publicize

the NAEP findings by improving the accessibility of the reports and engineering opportunities to bring NAEP into the public discourse. To this end, outside designers were employed to create attractive reports, an effort was made to make the prose understandable to laypersons, and 25,000 copies of each report were distributed nationally. Press conferences were held at the National Press Club in Washington, D.C., with appearances and commentaries by a variety of educational and policy personalities (e.g., notables such as Carl Sagan) and officials from teacher unions and subject area organizations. As a consequence of these actions, NAEP results were discussed on the news, in TV education specials, and on the “Today” show.

1984 to 2002: The Trial State Assessments and the Creation of the National Assessment Governing Board

Simultaneously with ETS’s NAEP activities, the foundation was being laid for the NAEP state-by-state assessments. The Education Amendments of 1984 extended NAEP for 5 years (with a budget of no less than \$4 million a year), **allowed state-by-state comparisons** [emphasis added], expanded the membership of APC, and appointed Assistant Secretary Chester Finn to represent the federal government on the latter.

Building on the positive public reaction to *A Nation at Risk*, in 1984, Secretary of Education Terrel H. Bell presented the idea of a large wall chart to display the comparative educational progress of each state. However, ED’s decision to use ACT or SAT scores—which were based on college-bound students—to populate this chart was widely criticized for not being an accurate representation of state results. All of this served to raise awareness that there *was* no appropriate state-level information for such a chart, and encouraged states to consider providing their own data.

One of the major leaders in the movement for state-level assessments was the Southern Regional Educational Board (SREB). In his “Notes on State NAEP,” prepared for the National Assessment Governing Board (Governing Board) 20th Anniversary Conference, Mark Musick (2009), president emeritus of SREB, explained:

In the late 1980s, eight SREB states pioneered a State NAEP pilot program that contained nearly all of the elements of what a few years later became the State NAEP Trial Assessment. Southern governors (the first three to step forward being Lamar Alexander [TN], Bob Graham [FL], and Chuck Robb [VA]), legislators, and educators had a strategy that focusing on progress, perhaps relatively surprising progress compared to the rest of the country, was an important strategy to change overall achievement and their state’s relative standing. As summed up by Governor Lamar Alexander, “It’s not just about where you are, but the direction in which you are moving” (Musick, 2009, p. 2).

Sentiment toward State NAEP was shifting, but, as described in *Overseeing the Nation’s Report Card* (Vinovskis, 1998), disagreement still existed about reporting state-level NAEP results. In 1984, when the Council of Chief State School Officers (CCSSO) voted on whether there should be a state-by-state NAEP, it passed by only one vote (20–19).

In 1986, then Secretary of Education William J. Bennett formed a distinguished 22-member NAEP study group, headed by Governor Lamar Alexander and H. Thomas James (former president of the Spencer Foundation), to study this issue. The Alexander-James report’s

most important recommendation was to provide expanded information for the whole country, and for every state within it (down to the community and neighborhood levels). The study group also made other important recommendations, including more concentration on the transition grades of 4, 8, and 12; expansion of the collection of background and school variables; and creation of the Educational Assessment Council (which later became the Governing Board) to oversee the redesign of NAEP. Consistent with the Alexander-James report recommendations, CCSSO, in 1987, established a National Assessment Planning Project to design a trial state assessment of mathematics at Grade 8 under a grant from the National Science Foundation (NSF) and USOE.

In 1998, Congress enacted P.L. 100-297, which had a substantial section devoted to NAEP (National Assessment of Educational Progress Improvement Act of 1998, SEC. 3402). This law—a precursor to the 2002 legislation still in effect—began with a statement of purpose as follows:

The purpose of this part is to improve the effectiveness of our Nation's schools by making objective information about student performance in selected learning areas available to policymakers at the national, regional, state, and local levels. To enhance its utility, such information shall be both representative and comparable and shall be maintained in a manner that ensures the privacy of individual students and their families.

This new NAEP law affirmed the placement of NAEP in NCES; set the assessment targets as the subjects of “reading, mathematics, science, writing, history/geography, and other areas” as well as adult literacy (the latter dependent on availability of appropriations); specified a trial state assessment (TSA) at the eighth grade in 1990 and a TSA at the fourth and eighth grades in 1992; called for an evaluation of the 1990 and 1992 TSAs by the National Academy of Sciences or the National Academy of Education; and directed the NCES commissioner to provide for continuing reviews of NAEP, including validation studies.

This legislation also established the Governing Board and laid out its responsibilities, which included selecting the areas to be assessed, detailing achievement goals, developing objectives, laying out test specifications, and having the final authority on the appropriateness of cognitive items. In December 1988, outgoing Secretary Bennett appointed the 23 Governing Board members, with Chester Finn as chair.

The next years were complicated, with NCES, the Governing Board, and CCSSO involved in decisions about the future of NAEP, and ETS and Westat, Inc., continuing to conduct national assessments as well as new state assessments. In 1990, the first TSA was conducted in mathematics at Grade 8 in 37 states and the District of Columbia; in 1992, the second TSA was conducted at Grades 4 and 8 in mathematics and Grade 4 in reading in 41 states, the District of Columbia, and two territories. Also, in 1992, five districts participated in NAEP's first federally funded Trial Urban District Assessment (TUDA).

In conjunction with the TSAs, the Governing Board implemented the process for establishing and reporting the results according to achievement levels. The NAEP achievement levels, which build on the ETS strategy of describing performance at various points on the NAEP achievement scales, are based on collective **judgments about what students should know and be able to do** [emphasis added] at each of three levels (*basic*,

proficient, and *advanced*),³ in each subject area assessment. According to the Governing Board's initial policy statement, this was to make the assessment "more useful to parents and policy makers as measures of performance in American schools and perhaps as an inducement to higher achievement" (National Assessment Governing Board, 1995, p. 1). Achievement levels were effective in increasing the visibility of NAEP but were broadly criticized (in evaluations by Governing Board consultants in 1991, the GAO in 1993, and a National Academy of Education Panel in 1993) for being too high and technically flawed.

In 1998, the nation's governors met in Charlottesville, Virginia, and committed themselves to a nationwide effort to reform education based on a core set of aspirations. They developed and agreed on six goals for improving the education system that guided their efforts (National Governors Association, 1998). In 1991, President George H.W. Bush released *America 2000: An Education Strategy*, which was based on the six national education goals. Goal 3 called for all students to leave Grades 4, 8, and 12 having demonstrated competency in challenging subject matter, and proposed the use of national tests at Grades 4, 8, and 12 to measure progress toward this goal (Alexander, 1991). In 1994, the Improving America's Schools Act (P.L. 103-384) included the six goals and stated that NAEP should "provide a fair and accurate presentation of education in reading, writing, and other subjects included in the third National Education Goal, regarding student achievement and citizenship."

From 1994 to 2002, national assessments were conducted every year (except 1999) and state assessments every other year. NCES was working to increase NAEP's use through a Secondary Analysis Grant Program (1992–2002) and the development of Web-based analysis tools that were released in 2001. The Internet was being recognized as a powerful way to disseminate NAEP results.

The National Academy of Education (NAE) was commissioned to evaluate the two TSAs (Glaser, Linn, & Bohrnstedt, 1997). Together with the American Institutes for Research (AIR), the NAE panel conducted a 6-year, ongoing review focused on the feasibility, validity, and utility of the TSAs, producing five reports that also covered the 1994 TSA. Consistent with recommendations in the early reports to maintain and expand the NAEP state assessments, the final NAE report in 1997 observed that the NAEP program had expanded the number of assessed students approximately fourfold in less than 10 years (1989–96). First and foremost, NAE reported, State NAEP had become an integral and welcome part of the NAEP program. Compared with the little-known NAEP that existed prior to the 1990s, the state-by-state assessments had led to increased attention on NAEP by a "host of interested parties, from the nation's capital to state assemblies and local school districts." Also a plus, NAEP had sought to include greater numbers of students with disabilities and students with limited English proficiency. The NAE report, however, did describe a strain on NAEP from the many demands being placed on it.

During this time, NAEP was receiving attention from the state results and, in particular, from reporting the nation's (and states') poor performance in reference to the achievement levels. It also was up for reauthorization in 1998, all of which resulted in numerous evaluations of NAEP beyond the NAE reports. Many of these reviews are summarized in

³ Pursuant to the updated Governing Board policy on achievement levels, adopted in November 2018, the names of the three achievement levels have been changed to *NAEP basic*, *NAEP proficient*, and *NAEP advanced*.

Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress (Pellegrino, Jones, & Mitchell, 1999).

According to the executive summary of *Grading the Nation's Report Card*, the committee focused its efforts on studying and improving the utility of the NAEP assessment results. In particular, the committee examined the use of NAEP data following the release of the 1996 mathematics and science assessments. The analysis relied on reports in popular and professional press, NAEP publications, and various letters, memoranda, and other unpublished documents. They determined that NAEP results were being used to:

- Describe the status of the education system.
- Describe the performance of students in different demographic groups.
- Identify the knowledge and skills over which students have (or do not have) mastery.
- Support judgments about the adequacy of observed performance.
- Argue the success or failure of instructional content and strategies.
- Discuss relationships among achievement and school and family variables.
- Reinforce the call for high academic standards and education reform.
- Argue for system and school accountability.

The committee further reported that NAEP was an important and useful monitor of students' academic achievement. It was doing a good job of meeting the descriptive needs of its users; people liked the kinds of information provided by the NAEP summary scores and achievement-level results. However, users of NAEP data also wanted and needed more information that would help them know what actions to take in response to NAEP results. The committee made two recommendations:

- Obtain more information about strengths and weaknesses in the knowledge and skills assessed by NAEP through in-depth analyses of student responses to NAEP items.
- Include NAEP in an integrated system of data collection so that information can be available about system-, school-, and student-level factors that relate to student achievement.

The first recommendation was manageable within NAEP, but the second went far beyond NAEP's mandate or influence. In support of the second recommendation, the committee also recommended strategies for transforming NAEP—streamlining it, but also using a multiple-methods design to capitalize on contemporary research and better assess achievement in such areas as scientific investigation, subject areas that not all students take, and difficult-to-assess populations (e.g., special-needs students and high school seniors).

2002: The Current NAEP Legislation

Understanding NAEP's responsibilities for providing data necessarily begins by examining the NAEP legislation. About 40 years after Tyler's recommendations to Keppel, in 2002, Congress enacted the National Assessment of Educational Progress Authorization Act (P.L. 107-279, which amended P.L. 107-110). This legislation, still in effect, elaborated considerably on the 1963 idea of what it means "to report annually on the progress of

students in the United States.” The best reference for understanding NAEP’s current roles and responsibilities, the legislation specifies that **NAEP has the responsibility to provide important policy-relevant information to help the United States improve school quality** [emphasis added]. This includes high-quality, accurate data about:

- Trends in educational achievement
- State-by-state trends in educational achievement
- Trends in improving equity in education
- Trends in improving excellence in education

In brief, the legislation is relatively detailed about the respective roles of the Governing Board and NCES in implementing NAEP. The legislation also is relatively prescriptive about what data NAEP should collect, that the data should be of high quality, how the data should be reported, and that NCES should report NAEP results on a regular basis.

Throughout the 2002 legislation, there is information about NAEP that defines its responsibilities and use. Defining the “National Assessment of Educational Progress” to collectively refer to national assessments, state assessments, and a long-term trend assessment in reading and mathematics, the legislation states that NAEP’s purpose “is to provide, in a timely manner, a fair and accurate measurement of student academic achievement and reporting of trends in reading, mathematics, and other subject matter (i.e., writing, history, geography, civics, economics, foreign languages, and arts).”

The legislation specifies that data should be collected in Grades 4, 8, and 12 in public and private schools. Furthermore, the reports should include information on special groups by region, race, ethnicity, socioeconomic status, gender, disability status, and limited English proficiency. Participation is voluntary for students, schools, and local education agencies. State participation, other than in reading and mathematics in Grades 4 and 8, also is voluntary.

The Governing Board, as part of its mandated responsibilities, is charged with developing assessment objectives and a process for reviewing the assessments that include the active participation of teachers, curriculum specialists, local school administrators, parents, and concerned members of the public. The Governing Board also is responsible for developing appropriate achievement levels for each grade or age in each subject area, while the Commissioner of Education Statistics is in charge of using such levels or other methods or indicators for reporting NAEP results.

The legislation further details what data NAEP should collect, the methods NAEP should use, and how the results should be reported. For example, reporting should include data trends at Grades 4 and 8 every 2 years, and, whenever feasible, the information collected should be cross-tabulated, compared, and reported by the special groups.

In summary, given the specified processes for program implementation, it is reasonable to conclude that the legislation intends for NAEP data to be used as a tool to monitor:

- National trends in students’ educational achievement at Grades 4, 8, and 12
- Trends in the percentage of students at each of the achievement levels set by the Governing Board

- State-by-state trends in reading and mathematics at Grades 4 and 8
- Educational achievement for groups of students by region, race, ethnicity, socioeconomic status, gender, disability status, and limited English proficiency

2002 to 2018: Assessments Conducted

According to the NCES website, NAEP currently assesses 10 subject areas. National assessments typically are conducted at Grades 4, 8, and 12, while state and TUDA assessments are administered at Grades 4 and 8. Except for 2002 and 2003, each of which had both national and state assessments, national-only assessments have been conducted every other year (even-numbered years), and combined national and state assessments have been conducted in the intervening (odd-numbered) years. Sample sizes for national-only assessments average 10,000–20,000 students per grade, while combined national and state assessments use samples of 3,000 students per grade per jurisdiction. Typically, 45–55 jurisdictions participate.

In the 16 years from 2002 to 2017, an estimated 8 million students⁴ have been assessed by NAEP. In accordance with the legislation, **reading and mathematics** have been assessed by far the most frequently of the 10 subject areas. Also, during this time period, participation in State NAEP assessments in reading and mathematics received a boost from the No Child Left Behind Act (NCLB) of 2001. The latter included the following provision: Any state that wishes to receive a Title 1 grant must include in the state plan it submits to the Secretary of Education an assurance that beginning in the 2002–2003 school year the state will participate in the biennial state-level National Assessment of Educational Progress (NAEP) in reading and mathematics at grades 4 and 8 (NCES, 2005).

In 2002, reading was assessed nationally and in states and TUDAs; in 2003, the reading assessments were repeated (except for the Grade 12 national assessment), and mathematics was added. Following this, both subjects were assessed nationally (at all three grade levels) and in states and TUDAs (at Grades 4 and 8) every 2 years (i.e., 2005, 2007, ... 2017). There also were state assessments in these subjects at Grade 12 in 2009 and 2013, and “long term trend” mathematics and reading assessments of 9-, 13-, and 17-year-olds (in school) in 2004, 2008, and 2012.

Regarding the other eight subject areas, during this period, **science** has been assessed four times nationally and in states, and twice in TUDAs. (The national assessments covered Grades 4, 8, and 12, and the state and TUDA assessments included Grades 4 and 8, except for 2011, when the entire science assessment was only at Grade 8.) **Writing** has been assessed four times nationally (but not at all three grades) since 2002 and three times in states and TUDAs in 2002, 2007 (at Grades 8 and 12), and 2017. Additional subject areas were assessed only nationally: **civics and U.S. history** three times (in 2006, 2010, and 2014) at Grades 4, 8, and 12 and once at Grade 8 only (2018); **geography** twice (in 2010 and 2014) at Grades 4, 8, and 12 and once at Grade 8 only (2018); **economics** twice (in 2006 and 2012)

⁴ There were nine state assessments that included nearly 7 million students (approximately 2,700 students in each of two grades and two subjects times about 50 jurisdictions = about 4.9 million, plus 1 million included in add-ons, such as Grade 12 or a third subject, and another 1 million included in from 5 to 27 TUDA assessments). There also were 10 national-only assessments, typically at three grades, and various special studies, which included another 1 million students or so. (personal communication from Keith Rust, Westat Inc.)

at Grade 12; **arts** twice (in 2008 and 2016) at Grade 8; and **technology and engineering literacy** twice (in 2014 and 2018) at Grade 8.

2002 to 2018: NAEP Report Cards Provided by NCES

As documented on the nationsreportcard.gov website maintained by NCES, the results of each of the approximately 40 NAEP assessments conducted over the past 15 years (see above) have been presented in a report card prepared by NCES by its contractors, principally ETS. A look across recent NCES reports of NAEP results, each with the Nation's Report Card logo, and all accessible online,⁵ reveals that the reports conform to the purposes, contents, and timeliness described in the legislation. As an example, the report of *NAEP 2015 Science* (NCES, 2016), released in October 2016, explains that the data are intended to “present a broad view of what our nation's students know and can do in science.” The website is in two major parts, one focusing on national results and the other on state results.

The national report begins with statistically significant changes in average science scores between 2009 and 2015 for Grades 4, 8, and 12. From there, the user can obtain (1) changes in average scores for groups defined by race/ethnicity, gender, national school lunch program, school location, region of country, and disability status; (2) changes in score gaps for racial/ethnic groups overall and by three science content areas (life science, physical science, and earth and space sciences); (3) changes in percentiles, shown in relation to the achievement-level cut scores; and (4) changes in average scores in the three science content areas. There are examples of skills demonstrated by students in answering questions.

The state report contains four major types of state-by-state results for Grades 4 and 8. The user can obtain (1) comparisons between states/jurisdictions and the nation in percentages of public school students performing at or above the *proficient* achievement level, (2) changes in average science scores between 2009 and 2015, (3) state score gaps between White and Black students, and (4) state-to-state score comparisons.

In summary, the *NAEP 2015 Science* report contains timely data on trends in achievement at Grades 4, 8, and 12 for the nation as well as for special groups; comparable data for states are available at Grades 4 and 8. The report closely follows the NAEP legislation by reporting all the different types of data that the legislation says are to be collected:

- National and state-level measures of academic achievement and data trends
- National and state-level performance according to achievement levels
- Cross-tabulations and comparisons of national and state-level achievement by special groups

From the *NAEP 2015 Science* report and other recent NCES reports, it seems that NAEP results are intended to be used for four major purposes:

- Monitoring trends in student educational achievement
- Determining if students are meeting appropriate levels of achievement

⁵ Recent reports, particularly those published in 2017, are Web-based and interactive.

- Tracking the progress of various subgroups (e.g., in accordance with NCLB, Every Student Succeeds Act)
- Making state-by-state comparisons

One goal of reporting is to facilitate the use of state-level results by the states themselves, and some states have used these data extensively. For example, when the 2005 results of the NAEP reading assessment revealed that eighth-grade students in North Carolina scored below the national average, the state deployed more than 200 literacy coaches around the state to help teachers reach students with reading difficulties.

With regard to reporting by achievement levels, the Institute of Education Sciences (IES) sponsored the recent *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress* (2017). The National Academies of Sciences, Engineering, and Medicine established the committee that conducted the evaluation. As part of its work, the committee conducted a broad Web-based search for references to NAEP results, and held a forum with panelists representing state and local policymakers, education writers, policy researchers, and advocacy groups as well as the designers, researchers, and technical experts who work on NAEP. They reported the uses of NAEP achievement-level data to be consistent with the results provided in the *Report Cards*:

- Trends or comparisons in proficiencies of successive cohorts
- Comparison with the results on the state assessment
- Point-in-time comparisons across states, districts, or population groups
- Ranking states or districts

In the section on use and interpretation (Chapter 6), the committee thus concluded that the NAEP achievement levels are disseminated and used by many audiences. However, they noted that interpretive guidance is lacking, and that the achievement-level descriptions may not provide enough information about what students at a given level know and can do.

In summary, the NAEP *Report Cards* are the vehicle for carrying out the legislation. They are produced in a timely manner and are easily accessible. One idea that has been deemphasized, however, is that of periodically reporting progress based on all the subject areas assessed. To complement the *Report Cards*' concentration on recent progress (current assessment compared with previous assessments) and percentages of students reaching the *NAEP proficient* and *NAEP advanced* achievement levels, NAEP could place a broader emphasis on trends. For example, a special "Report Card" synthesizing our nation's progress over the past 25 years across various subject areas would most likely reveal different trends across areas and subpopulations, together with a picture of overall progress for our country. Such a report would be in the spirit of NAEP's original overarching purpose.

2009 to 2018: NAEP Transitions to Digital-Based Assessments

With its responsibility as the Nation's Report Card and with its many users, it is important for NAEP to continue to be a leader in measuring the important skills and concepts that form the core educational content of our country's schools. However, assessing a broad range of subject areas using the most innovative testing approaches is very resource-intensive and challenging.

NAEP received a wake-up call to improve its relevance with the advent of the Common Core State Standards and the launch of the Obama administration's Race to the Top. During 2008–09, the National Governors Association (NGA) and CCSSO worked on a state-led process to create college- and career-ready standards that also involved leading education researchers, teachers, and extensive reviews by the public (Common Core State Standards Initiative, n.d.). More than 40 states and the District of Columbia joined together to launch the Common Core. Also, during the same period, the administration announced the \$4.35 billion Race to the Top Fund that offered competitive grants to encourage state education reforms, including the adoption of better standards and assessments (U.S. Department of Education, 2009).

More importantly for NAEP, in 2010, there was a separate Race to the Top Assessment Program (U.S. Department of Education, n.d.). This \$350 million grant competition was to support the work of consortia of states to develop and implement common, high-quality assessments aligned with common college- and career-ready K–12 standards.

According to the ED website, the “grant competition was designed to fill an urgent need in the nation’s education. It seeks to provide valid and instructionally useful assessments that provide accurate information about what students know and can do.” The secretary awarded 4-year grants to the Partnership for the Assessment of Readiness for College and Careers (PARCC), which had 14 states involved, and the Smarter Balanced Assessment Consortium (Smarter Balanced), which was comprised of 22 states. PARCC and Smarter Balanced each were provided an additional grant award of \$15.9 million to focus on (1) developing gap analyses between current and new standards, including standards in various areas such as curriculum, professional development, and assessments; (2) **enhancing technology to be used in the assessment systems** [emphasis added]; and (3) supporting educator understanding and use of assessment results (U.S. Department of Education, 2009).

By 2011, 1 year after the Common Core State Standards had officially been released, 45 states plus the District of Columbia had signed onto the standards and joined one or both of the assessment consortia. However, as controversy arose over the standards, the number of states planning to use the new assessments decreased sharply to 20 in 2016 (Jochim & McGuinn, 2016). The attrition occurred for various reasons, including parents’ views about too much testing, the perception of federal control of education, and negative reactions to the use of assessment results for teacher evaluations and school accountability. Although participation declined, a 2016 article in *U.S. News* reported positively about the new PARCC and Smarter Balanced assessments, concluding that they were more rigorous, required more critical thinking, and were better aligned to what students were learning compared with the old state tests (Sargrad, 2016). Also, these computerized assessments had made significant progress in measuring achievement of students with disabilities and English learners (ELs) by including built-in features, such as pop-up glossaries, zoom tools, and highlighters.

On its side, NAEP had been exploring new testing methods and question types that capitalized on the growing use of technology in education since 2001 (NCES, 2018). In 2009, computer interactive tasks were used to assess science; in 2011, writing was assessed using computers and word processing technology; and in 2014, the Technology and Engineering Literacy (TEL) assessment included scenario-based tasks to assess the framework in more authentic and direct ways. Spurred by the accomplishments of PARCC and Smarter Balanced, in 2015, NAEP piloted mathematics and reading assessments on tablets with an attached

keyboard, a stylus, and earbuds. Some test questions included multimedia, such as audio and video. Among other benefits, and consistent with the experience of the consortia assessments, these new technologies are improving NAEP's ability to offer accommodations to students with diverse backgrounds, including students with disabilities and ELs. In 2016, NAEP piloted mathematics and reading assessments that included online tools (such as calculators). Scenario-based tasks that allow students to retrieve, gather, and report information just as they do in their everyday lives also were being developed, starting with science but also including reading and mathematics.

In 2017, NAEP completed a major step into the future, using its experience with the 2015 pilot to fully transition to digital-based assessments for reading and mathematics, assessing over half a million fourth- and eighth-graders across the country. Concurrently, the digital-based writing assessment was given a second time (using tablets rather than computers). Moving forward, NAEP will use a digital format for all subjects it assesses, as reflected in 2018, when NAEP fielded digital assessments of civics, geography, and U.S. history at Grade 8. Pilot studies in 2018 included digital assessments of science at Grades 4 and 8 as well as reading and mathematics at Grade 12.

ADDITIONAL USES OF NAEP

In accordance with the 2002 legislation and in fulfillment of its role as the Nation's Report Card, NAEP maintains an extremely ambitious ongoing program of conducting assessments and reporting the results for use by the nation, population subgroups, and the states. However, NAEP has numerous other avenues for using and reporting data.

Using NAEP to Improve State Proficiency Standards

As early as 1996–97, the state of North Carolina implemented a new accountability program (the ABCs program) influenced by NAEP. Subsequently, the North Carolina General Assembly passed legislation in 1997 that says the following:

The State Board of Education shall develop a plan to create rigorous student academic performance standards for kindergarten through eighth grade and student academic performance standards for courses in grades 9–12. The performance standards shall align, whenever possible, with the student academic performance standards developed by the National Assessment of Educational Progress (NAEP). The plan also shall include clear and understandable methods of reporting individual student academic performance to parents (1997-221, s. 3(e)).

Building on this need for understanding—and tightening—the connection between state standards, state assessments, and external benchmarks, NCES's periodic reports titled *Mapping State Proficiency Standards* (NCES, 2017a) have provided information that shows where each state's standards, as measured by the state's assessment, lie on the NAEP scale. Using data from each state assessment and from NAEP, NCES places each state's standard for *proficient* performance in reading and mathematics onto a common scale defined by NAEP scores. This process of "state mapping"—which has been carried out for every NAEP administration since 2003—allows each state to compare the stringency of its criteria with criteria used by other states. The results show wide differences across states, and many states have levels of *proficient* performance that fall within NAEP's *basic* level or lower.

Also, since 2005, researchers at *Education Next* have used NAEP data to grade state proficiency standards on an A–F scale by comparing the percentage of students identified by states as *proficient* in reading and mathematics with the percentage of *proficient* students reported by State NAEP (Hamlin & Peterson, 2018). From 2005 to 2009, on average, states made no progress toward raising their proficiency standards and reducing the gap with NAEP, and, in 2009, no state received an "A" for being aligned with NAEP. However, after the Common Core State Standards movement was initiated, the rigor of state proficiency standards increased sharply, especially between 2013 and 2015, and the gap with NAEP was narrowed substantially to 10 percentage points, on average. This narrowing of the gap was maintained in 2017, resulting in *Education Next* "A" grades for 16 states and the District of Columbia.

A number of states have used this type of information in significant ways. For example, the state of Tennessee made remarkable progress in the *Education Next* results—from an F in 2009 to an A in 2017—and was ranked at the top for improvement in standards. According to the online article "Tennessee's Education Transformation," Tennessee first committed to raising its education standards in 2007 after the U.S. Chamber of Commerce used the NAEP mapping results to give Tennessee an F for low academic expectations and lack of truth in advertising regarding the accomplishments of the state's K–12 education system (Tennessee

Office of the Governor, n.d.). Under then governor, Bill Haslam, Tennessee implemented new academic standards and launched an assessment known as TNReady. These gains have been reflected in NAEP state achievement data, supporting the idea that Tennessee is on the path toward meeting its achievement goals.

NAEP also played a prominent role in Massachusetts' standard-setting decisions, both for the state assessment (Massachusetts Comprehensive Assessment System, or MCAS) and for the PARCC assessment developed when Massachusetts was part of that state consortium. According to Jeffrey Nellhaus, who has served as director for both programs, NAEP's achievement-level descriptors informed the development of the descriptors for each program (personal communication, 2018). First, language from the description of NAEP's *proficient* level was reflected in the definitions of *proficient* for MCAS and Level 4 for PARCC. (PARCC's Level 4 is called "Meets Expectations," for being on track or ready for college and careers). Second, NAEP results in reading and mathematics at Grades 4 and 8 were used as benchmarks by both the standard-setting panels charged with recommending cut scores for the PARCC assessment and the chief state school officers who finalized the cut scores. There was high interest among the PARCC states in having the initial results of the PARCC assessment be consistent with NAEP.

NAEP Special Studies

Over the years, NAEP has conducted a number of special studies. For example, it periodically conducts and reports on the 12th-grade High School Transcript Study. Example data from the transcript study include gender differences in science, technology, engineering, and mathematics (STEM) interests and relationships between grade point average and NAEP achievement. Another special study, the National Indian Education Study, is designed to describe the condition of education for American Indian and Alaska Native students. Conducted in 2005, 2007, 2009, 2011, and 2015, the study provides information about academic performance in reading and mathematics at the fourth and eighth grades as well as students' self-reported exposure to Native American culture.

NAEP also coordinates special studies that involve special data collection procedures or samples. These include the Oral Reading Study (2002), the Charter School Pilot Study that reported performance in charter schools (2003), a trend report called *Student Achievement in Private Schools* (2005), *Mathematics in Puerto Rico* (2015), and the Technology-Based Assessment Project, which looked at the potential of assessing mathematics, writing, and problem solving by computer (2016). Other special studies are based on secondary analyses of NAEP data—for example, *Measuring Status and Change in NAEP Inclusion Rates of Students with Disabilities*, which analyzed trends between 2007 and 2009, and *Achievement Gaps*, which looked at trends in differences in achievement between Black and White students and between Hispanic and White students (2015).

Some of the most interesting and policy-relevant special studies involve connecting NAEP to other data. One example is the *Mapping State Proficiency Standards* reports described earlier; another is the NAEP-TIMSS linking study, which was based on 2011 results at the eighth grade and linked the NAEP scale to the Trends in International Mathematics and Science Study (TIMSS) scale, so states could compare the performance of their students with that of students in other countries (Jia et al., 2014).

NAEP Data Explorer and Availability of Data for Secondary Analysis

NAEP encourages researchers and policymakers to make use of its rich source of information, and has developed the NAEP Data Explorer (NDE) to encourage and greatly facilitate the use and secondary analysis of NAEP results. For example, visitors to the “Report Card” website also are invited to use the Data Explorer to view achievement results for the nation, by state, and in relation to contextual information collected using student, teacher, and school questionnaires that are administered in conjunction with the assessments.

As explained on the ETS website, NDE is an interactive Web tool that facilitates (1) building custom statistical tables, animated graphics, and state maps; (2) exploring decades of assessment results for various subjects, grades, and jurisdictions; and (3) viewing information about factors that may be related to student learning. NCES’s website explains that most NAEP data from 1990 to date are available in NDE for public use.

NCES routinely provides events and materials on how to analyze NAEP data and a yearly training in using the NAEP database; in addition, micro-level NAEP data in raw format (i.e., respondent-level data, including weights) are available upon receipt of a restricted-use data license from NCES. IES invites applications to its annual grant program for studies using NAEP data, and offers research funding webinars to provide guidance for potential applicants, while small grants for dissertations and research using NAEP also are available through the training and research program administered by the American Educational Research Association (AERA).

A recent project by Reardon, Kalogrides, and Ho (2016) offers an example of work using NAEP micro-data. Working under grants from IES, the Spencer Foundation, and the William T. Grant Foundation, these researchers created a database linking each state’s assessment scale to the NAEP scale and applying the transformation to a database of district-level means. This important research extends the uses of NAEP by enabling large-scale education research about national variation in district achievement (see Reardon, Kalogrides, and Ho, 2016).

Availability of Frameworks and Questions

In addition to the data, NAEP has other tools and applications useful for educational improvement at the local level. For example, the NAEP frameworks in 11 subject areas, each containing the blueprints for the content and design of the associated assessment, are available on the Governing Board’s website (nagb.gov). For each framework, the Governing Board works with a committee of subject-matter experts, practitioners, and members of the general public (e.g., researchers and policymakers) to develop a set of standards for what students should know and be able to do. According to the Governing Board website, without advocating for any particular approach to instruction, NAEP frameworks provide a starting point for constructive conversations about high-quality educational standards and assessments.

Also, NCES provides the “NAEP Questions Tool,” which enables (1) examination of the many released items from previous NAEP assessments by subject and grade, (2) an opportunity to test yourself on NAEP items, and (3) downloading of preselected or customized assessments by subject or grade. (The quality of some of the items and the user-friendliness of the tools, however, may depend on the circumstances.) There also are “Item Maps” showing the relationship between the NAEP achievement scales and some of the items, and a special part of the website presents the interactive computer tasks (ICTs) from

the 2009 science assessment. As another resource, the framework for contextual data collection is available on the Governing Board website, and the Contextual Questionnaires are available on the NCES website.

Montana has taken advantage of these materials to develop several online courses for teachers (opi.mt.gov/Leadership/Assessment). As described on the website, “PAO (Process, Assessment, Outcome) Instruction” includes courses in English language arts, math, and science that provide hands-on experience with item classification by allowing teachers to align high-quality NAEP-released test items to the Montana Common Core Standards. There also is a course on ICTs that familiarizes teachers with the ICTs developed for NAEP’s 2009 science assessments. With regard to this course, the website explains that as Montana progresses toward computer-based testing, teachers can use ICTs in the classroom to better prepare students. This course provides a tutorial on using ICTs with a Grade 6–8 science classroom as a model.

In general, the accessibility of the NAEP frameworks and items help states and districts that are designing their own standards and assessments. In the Montana example, courses for teachers were designed to help them understand standards, how items relate to standards, and features of computer-based assessments. In summary, Montana appears to have found ways to use NAEP to support learning about assessments, but there does not appear to be state support for using the NAEP frameworks and items to improve curriculum and instruction—something that was implemented in early NAEP and recommended in the 1999 *Grading the Nation’s Report Card* (Pellegrino et al., 1999). To fill this gap, NAEP might encourage another organization (e.g., NSF) to support an analysis of students’ responses to items in the most recent assessment of a particular subject area (e.g., science), with the goal of informing policymakers, educators, and teachers about how best to improve curriculum and instruction in that area.

National Assessment Governing Board Studies and Reports

Aside from producing the NAEP assessment frameworks, the Governing Board has authored or commissioned numerous reports on topics related to the Nation’s Report Card and the Board’s work. In the past 5 years (2012–17), it has prepared 18 publications covering the topics of students’ academic preparedness, achievement levels, contextual data, science, outreach, and Board operations. The Governing Board is particularly interested in validity evidence for statements in NAEP reports about 12th-grade students’ academic preparedness for college. For example, they conducted a survey of a nationally representative sample of postsecondary institutions regarding the tests and cut scores used to determine whether entry-level students will need remedial/developmental instruction in reading or mathematics. The most widely used college admissions tests (e.g., ACT and SAT) also have been compared with NAEP through statistical linking studies (linking NAEP scores to scores on other tests and to follow-up data for students who took NAEP) that were conducted to place college readiness benchmarks and other outcome data on the NAEP scale. More recently, the Governing Board has explored the potential for deriving indicators of postsecondary preparedness from NAEP assessments and questionnaires.

Using NAEP Data to Improve NAEP

The NAEP legislation specifies that the “data should be of high quality” and provide “fair and accurate measurement” of students’ academic achievement in reading, mathematics, and

other subject matter (i.e., writing, history, geography, civics, economics, foreign languages, and the arts). Consistent with this mandate, considerable energy is devoted to research on how to improve the validity and accuracy of NAEP procedures and results.

Since 1995, AIR has maintained the NAEP Validity Studies (NVS) Panel,⁶ an independent panel of experts that meets to commission and discuss research addressing validity considerations for NAEP. In 2014, AIR developed a comprehensive NAEP validity research framework for the Panel and reviewed relevant research to produce a database of 180 unique studies related to topics in the validity framework. Also, according to the AIR website, the NVS Panel has produced 34 reports evaluating NAEP procedures and providing a more in-depth look at NAEP results. For example, reports published during the past 5 years have addressed:

- Alternative approaches to setting NAEP performance standards, NAEP achievement in relation to school testing conditions, and validity issues involved in making cross-grade statements about NAEP results (2012)
- Validity of the NAEP full population estimates, implications for NAEP of the Common Core State Standards, and feasibility of an accessible block alternative in mathematics (2013)
- Linking the NAEP scale in reading to the Progress in International Reading Literacy Study (PIRLS) scale (2014)
- Maintaining NAEP's validity in a Common Core-based environment and alignment of NAEP 2015 mathematics items with the Common Core (2015)
- Identifying the relationships between task-text-reader interactions and item difficulty (with the goal of improving the assessment of reading comprehension (2017)
- Developing new indices to measure digital technology access and familiarity (2018)

In recognition of the importance of the transition to digital-based assessments, the NVS Panel is currently conducting or planning research to examine the impact of visual and interactive features in assessment, NAEP's capacity for measuring skills learned by students who are using "cutting-edge" mathematics curricula, and implications for trend reporting in an era of rapidly changing digital devices and interfaces. Understanding innovative instructional approaches and the capabilities of digital-based assessments represent an exciting future for NAEP.

The Design and Analysis Committee (DAC), which has its roots in the original NAEP TAC and ANAC, has been maintained by ETS since 1983. Among many technical areas, DAC considers sampling, statistical, and psychometric aspects of NAEP operations, analysis, and reporting.

NAEP's Potential to Inform Policy

Taken together, the *Report Cards* and the many NAEP special studies, especially those linking NAEP to other databases ranging from district to international levels, produce an enormous amount of information about policy issues central to our nation's future in providing equitable and high-quality education.

⁶ This white paper is an NVS Panel project being prepared under the guidance of the Panel, AIR, and NCES.

NAEP has influenced equity and social justice in education through the vast amount of data it has produced about gaps in achievement between demographic subgroups as well as for special populations, such as students attending schools in urban districts and through the Indian Education Study. The questionnaires given to students, teachers, and school administrators collect additional information that helps put student achievement results in context and allows meaningful comparison between student groups (NCES, 2018). By asking questions about students' opportunity to learn as well as their learning experiences, NAEP questionnaires provide important information for educators, policymakers, and researchers that allows in-depth analyses to better understand the context in which students learn. To facilitate their use, the results of these contextual questionnaires are sorted into eight broad categories in the NAEP Data Explorer: Major Reporting Groups, Student Factors, Factors Beyond School, Instructional Content and Practice, Teacher Factors, School Factors, Community Factors, and Government Factors.

Information about excellence in education has been enhanced through the development and use of NAEP achievement levels and through the Governing Board's research about college and career readiness, which includes linking NAEP outcomes to proficiency on college placement tests. In addition, the NAEP state mapping studies have served to raise states' standards.

NAEP provides relevant information that is consistent over time for an ever-increasing range of constituents. The state-level and TUDA assessments have been operational since 2002, and, more recently, NAEP has been linked to district-level data (see Reardon, Kalogrides, and Ho's 2016 report, *Linking U.S. School District Test Score Distributions to a Common Scale*). To provide a comparison with achievement in other countries, NAEP also presents states with an international context for their results by linking U.S. state-by-state achievement to the TIMSS and PIRLS assessments of mathematics and reading in approximately 60 countries (see the NAEP-TIMSS linking study and NAEP-PIRLS linking study).

All in all, NAEP has the following important policy-relevant uses:

- National monitoring (Is the quality of education in our country improving?)
- State-by-state monitoring in a national context (and a global context through linking)
- Providing a model for state assessments (e.g., item formats, achievement levels)
- Providing states with additional arguments for new education funding
- Promoting equity through regular monitoring of trends in achievement gaps by race, ethnicity, socioeconomic status, gender, disability status, and limited English proficiency
- Promoting equity by monitoring trends in achievement for urban school districts and through the Indian Education Study
- Promoting excellence through reporting levels of student achievement
- Promoting excellence through research on how to use NAEP to provide indicators of postsecondary preparedness

Next Steps for NAEP

It is clear that NAEP has come a long way from Ralph Tyler's original memorandum to become an enormous, valuable, and longstanding enterprise with important monitoring and policy uses for the United States. By addressing issues of educational equity by gender and among various demographic groups, NAEP identifies which groups of students need particular support if the country is to effect upward change. In reporting trends in levels of educational achievement, it highlights the quality and progress (successes and failures) of our education system in core subject areas. NAEP does work vital to maintaining the educational health of our nation and keeping the United States on a course of educational improvement.

In its 50 years of operation, NAEP has met a wide range of national needs by greatly expanding the network of constituents to which it is relevant as well as by increasing its relevance to states and districts. NAEP data are well positioned to be in the forefront of education policy discussions about how the United States is faring educationally in a global context as well as the differential impact of individual states on the national standing.

NAEP is to be highly commended for the substantial effort it has invested in improving the validity and accuracy of assessment procedures and data over its history, especially more recently. Finally, the monumental amount of NAEP data and tools to support data use available on the NCES, Governing Board, ETS, AIR, and other websites represents a true national treasure that needs to be more thoughtfully and thoroughly mined.

Despite NAEP's prodigious successes, in some ways it perpetually faces the same major issues of quality, relevance, and costs decade after decade:

- Fulfilling the requirements of the NAEP legislation by routinely providing reliable data
- Remaining in the forefront of excellence and innovation in large-scale assessment
- Improving the assessment content coverage to better reflect what students are taught in school (including today's difficult-to-measure subject areas and 21st century skills)
- Increasing the efficiency of the assessments to reduce costs
- Developing successful reporting and dissemination strategies to increase NAEP's use

Both NCES and the Governing Board have strategic plans for the coming years of NAEP. The NCES strategic plan, developed in 2017, has five goals and 11 directions to support these goals (NCES, 2017b). The first two goals are intended to keep NAEP as the most authoritative source of information concerning academic achievement of American youth by (1) strengthening the utility of NAEP data through linkages, alignments, and additional reporting variables; and (2) adopting new approaches to assessment content that are responsive to changes in the educational environment. The last three goals support the vision of NAEP as a model of excellence and innovation in large-scale assessment by (3) leading in assessment innovation through research and technology, (4) improving the usability of NAEP data through enhanced reporting and by providing resources for best practices and lessons learned, and (5) improving operational processes by increasing efficiency through technology, modifying NAEP's design, and integrating components with other NCES assessments. The 11 directions are grouped into two categories—current directions and future directions. The current directions reflect important initiatives to ensure

that NAEP remains a leader in efficiency, validity, and relevance. The future directions, which reflect NCES' priorities for what NAEP should become, emphasize:

- Greater student interactivity
- Reduced student burden
- Increased usefulness of reporting for teachers and parents
- Sophisticated use of technology to assess new knowledge and skills

The Governing Board unanimously approved a strategic vision in 2016 that is intended to focus the Board's work through 2020. The vision was developed to expand NAEP's impact by addressing the question: "How can NAEP provide information about how our students are doing in the most innovative, informative, and impactful ways?" (National Assessment Governing Board, 2016). As summarized below, the strategic vision has a dual focus of (1) informing stakeholders to expand NAEP's use, and (2) innovating to enhance NAEP's form and content.

- Inform—Facilitate awareness and use of NAEP resources (e.g., results, measurement innovations, frameworks, contextual variables) by (1) strengthening partnerships; (2) increasing connections between NAEP data, administrative data, and state, national, and international student assessments; (3) expanding the availability of NAEP resources and creating new resources; and (4) promoting sustained dissemination and use of NAEP information beyond *Report Card* releases.
- Innovate—Use technology to revise NAEP's design, form, and content to keep NAEP in the forefront of measuring and reporting student achievement by (1) updating frameworks while maintaining student achievement trends; (2) improving the content, analysis, and reporting of NAEP's contextual variables; (3) researching implications related to the future of NAEP's long-term trend assessments; (4) researching assessments used in other countries; (5) developing policy approaches to revising the assessment subjects and schedule based on the nation's evolving needs; and (6) developing new approaches to measure the complex skills required for transition to postsecondary education and career.

As featured prominently in both strategic plans, technology provides an avenue for improving NAEP's capacity to support in-depth understanding of student achievement by assessing an increased variety of subject areas and more challenging educational goals. Even though increased use of technology has been recommended in previous studies of NAEP (e.g., NAE studies), the actual use of technology as a basis for assessment was not feasible until recently. Now that NAEP has transitioned to digital-based assessments, it will likely require considerable NAEP exploration and research to maximize the measurement potential of this innovation. That is, NAEP needs to systematically move toward assessing a widely expanded view of achievement that will meet the needs of our 21st century society. The move to digital-based assessments also has considerable potential to develop more efficient data collection and measurement methods.

In contrast to the promise that technology provides for improving assessment content and form, technology is not a panacea for addressing the challenge, which has persisted across NAEP's entire history, of making NAEP more useful and well known. Although promoting the use of NAEP results is central to current NCES and Governing Board

strategic plans and visions, supporting a more widely used and better recognized NAEP also has been included in NAEP goals and NAEP evaluation reports since NAEP's beginning 50 years ago. Technology may facilitate dissemination of information, but the fact that this persists as a goal provides evidence that making NAEP more useful and well known remains extremely challenging. The ideas and resolution reflected in the NCES and Governing Board strategic plans and visions, however, offer hope for the future of a highly regarded and well-recognized NAEP.

REFERENCES

- Alexander, L. (1991). *America 2000: An education strategy sourcebook*. Washington, DC: U.S. Department of Education.
- Beaton, A. E., Rogers, A. M., Gonzalez, E., Hanly, M. B., Kolstad, A., Rust, K. F., ... Jia, Y. (2011). *The NAEP primer, chapter 2.1* (NCES 2011-463). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved from <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011463>
- Common Core State Standards Initiative. (n.d.). *Development process*. Retrieved from <http://www.corestandards.org/about-the-standards/development-process/>
- Education Commission of the States. (n.d.). *History*. Retrieved from <https://www.ecs.org/about-us/history/>
- Gardner, J. (2004). An interview with John Gardner. In L. V. Jones & I. Olkin (Eds.), *The Nation's Report Card: Evolution and perspectives* (pp. 113–121). Bloomington, IN: Phi Delta Kappa Educational Foundation in cooperation with the American Educational Research Association.
- Gilford, D. M. (2004). NAEP and the U.S Office of Education, 1971 to 1974. In L. V. Jones & I. Olkin (Eds.), *The Nation's Report Card: Evolution and perspectives* (pp. 165–183). Bloomington, IN: Phi Delta Kappa Educational Foundation in cooperation with the American Educational Research Association.
- Glaser, R., Linn, R., & Bohrnstedt, G. (1997). *Assessment in transition: Monitoring the nation's educational progress*. Stanford, CA: National Academy of Education.
- Hamlin, D., & Peterson, P. E. (2018). Have states maintained high expectations for student performance? *Education Next*, 18(4).
- Hazlett, J. A. (1974). *A history of the National Assessment of Educational Progress, 1963-1973* (unpublished doctoral dissertation). University of Kansas, Lawrence, KS.
- Jia, Y., Phillips, G., Wise, L., Rahman, T., Xu, X., Wiley, C., & Diaz, T. E. (2014). *2011 NAEP-TIMSS linking study: Technical report on the linking methodologies and their evaluations* (NCES 2014-461). Retrieved from <https://nces.ed.gov/nationsreportcard/subject/publications/studies/pdf/2014461.pdf>
- Jochim, A., & McGuinn, P. (2016). The politics of the Common Core assessments. *Education Next*, 16(4).
- Jones, L. V. (2004). Chronology, 1963-2003. In L. V. Jones & I. Olkin (Eds.), *The Nation's Report Card: Evolution and perspectives* (pp. 11–23). Bloomington, IN: Phi Delta Kappa Educational Foundation in cooperation with the American Educational Research Association.

- Lapointe, A. (2004). NAEP and the U.S Office of Education, 1971 to 1974. In L. V. Jones & I. Olkin (Eds.), *The Nation's Report Card: Evolution and perspectives* (pp. 185–199). Bloomington, IN: Phi Delta Kappa Educational Foundation in cooperation with the American Educational Research Association.
- Messick, S., Beaton, A., & Lord, F. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era* (NAEP Report No. 83-1). Princeton, NJ: Educational Testing Service.
- Musick, M. (2009, March). *Notes on State NAEP*. Paper presented at the National Assessment Governing Board 20th Anniversary Conference, Washington, DC. Retrieved from <https://www.nagb.gov/content/nagb/assets/documents/who-we-are/20-anniversary/musick-state-NAEP.pdf>
- National Academies of Sciences, Engineering, and Medicine. (2017). *Evaluation of the achievement levels for mathematics and reading on the National Assessment of Educational Progress* (pp 189–218). Washington, DC: The National Academies Press.
- National Assessment Governing Board. (1995). *Developing student achievement levels for the National Assessment of Educational Progress: Policy statement*. Washington, DC: Author.
- National Assessment Governing Board. (2018). *Developing student achievement levels for the National Assessment of Educational Progress: Policy statement*. Washington, DC: Author. Retrieved from <https://www.nagb.gov/content/nagb/assets/documents/policies/ALS-revised-policy-statement-11-17-18.pdf>
- National Assessment Governing Board. (2016). *Strategic vision*. Washington, DC: Author. Retrieved from <https://www.nagb.gov/content/nagb/assets/documents/newsroom/press-releases/2016/nagb-strategic-vision.pdf>
- National Assessment of Educational Progress Authorization Act of 2002, Pub. L. 107-279, 20 USC 9501.
- National Assessment of Educational Progress Improvement Act of 1988, Pub. L. 100-297, 20 USC 3402.
- National Assessment of Educational Progress. (2014). *NAEP validity studies framework*. Washington, DC: American Institutes for Research.
- National Center for Education Statistics. (2005). *Important aspects of No Child Left Behind relevant to NAEP*. Retrieved from <https://nces.ed.gov/nationsreportcard/nclb.aspx>
- National Center for Education Statistics. (2016). *The nation's report card: 2015 science assessment*. Retrieved from https://www.nationsreportcard.gov/science_2015/#?grade=4
- National Center for Education Statistics. (2017a). *Mapping state proficiency standards*. Retrieved from <https://nces.ed.gov/nationsreportcard/studies/statemapping/>

- National Center for Education Statistics. (2017b). *NAEP strategic plan*.
- National Center for Education Statistics. (2018a). *NAEP digitally based assessments*. Retrieved from <https://nces.ed.gov/nationsreportcard/dba>
- National Center for Education Statistics. (2018b). *NAEP survey questionnaires—Questionnaires for students, teachers, and school administrators*. Retrieved from https://nces.ed.gov/nationsreportcard/experience/survey_questionnaires.aspx
- National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office. Retrieved from https://www.edreform.com/wp-content/uploads/2013/02/A_Nation_At_Risk_1983.pdf
- National Governors Association. (1998). *National indicator initiatives*. Washington, DC: Author. Retrieved from https://classic.nga.org/cms/home/nga-center-for-best-practices/center-publications/page-archive/col2-content/title_national-indicator-initiatives.html
- Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.) (1999). *Grading the Nation's Report Card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.
- Reardon, S. F., Kalogrides, D., & Ho, A. (2016). *Linking U.S. school district test score distributions to a common scale, 2009-2013* [CEPA Working Paper No. 16-09]. Retrieved from <https://cepa.stanford.edu/sites/default/files/wp16-09-v201604.pdf>
- Sargrad, S. (2016). Building better tests. *U.S. News*. Retrieved from <https://usnews.com/opinion/knowledge-bank/articles/2016-03-04/parcc-and-smarter-balanced-are-building-better-standardized-tests>
- Tennessee Office of the Governor. (n.d.). *Tennessee's education transformation*. Nashville, TN: Author. Retrieved from <https://www.tn.gov/former-governor-haslam/priorities/tennessee-education-transformation.html>
- U.S. Department of Education. (2009). *Race to the Top Fund*. Retrieved from <https://www2.ed.gov/programs/racetothetop/index.html>
- U.S. Department of Education. (n.d.). *Race to the Top Assessment Program*. Retrieved from <https://www.ed.gov/open/plan/race-top-assessment>
- U.S. General Accounting Office. (1976). *The National Assessment of Educational Progress: Its results need to be made more useful*. (HRD-76-113). Washington, DC: GAO.
- Vinovskis, M. A. (1998). *Overseeing the Nation's Report Card: The creation and evolution of the National Assessment Governing Board (NAGB)*. Ann Arbor, MI: University of Michigan. Retrieved from <https://www.nagb.gov/content/nagb/assets/documents/publications/95222.pdf>

References

Wirtz, W., & Lapointe, A. (1982). *Measuring the quality of education: A report on assessing educational progress*. Washington, DC: Authors.