The Influence of External Evaluations on the National Assessment of Educational Progress

CSE Technical Report 548

Robert L. Linn
CRESST / University of Colorado at Boulder

September 2001

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.5 Coherence and Collaboration: Equity and Technical & Functional Quality
Robert L. Linn, Project Director, CRESST / University of Colorado at Boulder

The work reported herein was partially supported by the National Institute of Statistical Sciences and partially under the Educational Research and Development Centers Program, PR Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect positions or polices of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education; nor do they necessarily reflect the positions or policies of the National Institute of Statistical Sciences.

THE INFLUENCE OF EXTERNAL EVALUATIONS ON THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Robert L. Linn

CRESST/University of Colorado at Boulder

Abstract

The National Assessment of Educational Progress (NAEP) has been the subject of many evaluations during its history. Those evaluations are reviewed, and the influences of recommendations made in the evaluations are discussed. It is concluded that many of the recommendations of evaluators have been heeded by those responsible for the design and conduct of NAEP.

The National Assessment of Educational Progress (NAEP) has been the subject of many evaluations, both formal and informal, during its history. Some of the evaluations have been mandated by Congress. Others have been conducted in response to requests from the National Center for Education Statistics (NCES) or the National Assessment Governing Board (NAGB). Still other evaluations have been undertaken at the initiative of individual scholars or have occurred as part of the deliberations of committees formed to provide advice to those responsible for NAEP.

The purpose of this report is to review the evaluations that have been conducted and the recommendations that have been offered and to see the extent to which the evaluations have influenced the way that NAEP is conducted and reported. For the most part, we will limit our review to external evaluations available in the literature. Inclusion of the evaluative conclusions and recommendations made over the years by the large number of NAEP advisory committees that have worked for NCES, NAGB, and the NAEP contractors would be unwieldy and often would suffer from incomplete documentation. Hence, advisory committee recommendations will generally be considered beyond the scope of this review. Exceptions to this rule are made when secondary sources refer to committee recommendations.

Initial Proposals for NAEP and Early Reactions

The idea for NAEP is generally traced to Francis Keppel. During his tenure as commissioner of education, Keppel was concerned that he did not have a good basis for fulfilling the responsibility of the Office of Education to report periodically to Congress on the progress of education in states that was part of the 1867 charter for the United States Office of Education (Fitzharris, 1993, p. 23). In 1963, Keppel asked Ralph Tyler to consider the feasibility of developing a periodic national assessment of student learning (Hazlett, 1974, p. 26). Tyler's oft-cited 1963 memorandum gave an endorsement to the idea of a national assessment and provided a broad outline for the creation of NAEP. With sponsorship by the Carnegie Foundation, three conferences were held where the idea of a national assessment was elaborated and reactions were obtained.

Tyler's influential role in shaping NAEP, articulating its purposes, and generating support for the idea was evident even at the first conference that was held in December 1963. As described by Greenbaum, Garet, and Solomon (1975), Tyler's presentation at that initial conference

> emphasized that (1) the Assessment would test **general** levels of knowledge, "what people have learned, not necessarily all within the school system," (2) the tests would not be aimed at discriminating among individuals, unlike most educational tests, (3) there would be an attempt to assess more accurately the levels of learning of the least educated, average, and most educated groups in the society, (4) some sort of matrix sampling system would test individuals only on a small number of questions but results could be aggregated to reflect the knowledge of particular subgroups in the population, (5) adults might be included in the sample, (6) stages, such as the end of elementary school, the end of intermediate school, and the end of high school, should be used in connection with specific testing ages rather than at specific grade levels, and (7) the effects of the tests themselves would have to be carefully considered because they might become standards for educational curricula and might also reflect on the status of particular communities. (p. 10, emphasis in the original)

These core ideas were elaborated with more detail, amended with some additions and refined during the following six years before the first assessment was administered, and some additions were made to the list (e.g., the encouragement of the use of short-answer items and performance tasks, the consensus process for determining learning objectives to be assessed, the use of NAEP personnel to administer assessments, and the reporting of results by exercise rather than on a

composite scale). The broad conception articulated by Tyler at the initial conference, however, remained remarkably intact as a conceptual blueprint for NAEP both prior to and throughout the years that NAEP was conducted by the Education Commission of the States (1969-1982).

Evaluative reactions began to be heard shortly after Tyler's memorandum became available and the Carnegie conferences were held. The idea for a national assessment was viewed with alarm by several organizations of educators (e.g., the National Education Association, the American Association of School Administrators [AASA]; Hazlett, 1974). Such groups saw the idea of a national assessment as a national test that would threaten local autonomy and introduce a national curriculum. Concerns about a national assessment that might introduce aspects of a national curriculum and undermine local control were allayed by the ways in which NAEP came to be defined and implemented. However, as we shall see, such concerns have resurfaced in periods when the role of NAEP has been expanded or when the expansion of its roles has been proposed.

One of the suggestions made at the second conference sponsored by the Carnegie Foundation to explore the creation of a national assessment was that reports be provided at the state level (Hazlett, 1974, p. 49). Although it is unclear how much influence reactions such as those of the AASA had in shaping NAEP, it is clear that the suggestion to have state-level sampling and reporting was eschewed in favor of a plan that provided results for the nation as a whole, for broad geographical regions, and for segments of the population defined by student demographic characteristics.

The Role and Functions of NAEP

The example of differences in views about the reporting of NAEP results at the state level is just one of many considerations that had to be resolved in defining the roles and functions of NAEP and in developing a system consistent with those roles and functions. An obvious role of NAEP is that of an indicator. Early discussions suggested that NAEP might produce indices analogous to the gross national product or the U.S. Consumer Price Index (Hazlett, 1974). Educational indicators prior to the development of NAEP had consisted largely of status characteristics such as enrollment, teacher qualifications, and expenditures.

The early defining characteristics of NAEP dealt more with approach and design than with purpose or use. At the most global level, there is widespread and

long-standing agreement that the purpose is to contribute to the improvement of education through the process of providing better information to policymakers, educators, and the public. Frank Womer, the staff director of NAEP at that time, made this clear in one of the NAEP publications, entitled *What is National Assessment?* According to Womer (1970): "The ultimate goal of National Assessment is to provide information that can be used to improve the educational process, to improve education at any and all of its levels where knowledge will be useful about what students know, what skills they have developed, or what their attitudes are" (p. 1).

Although sometimes only implicit, one of the fundamental ideas leading to the creation of NAEP is that information about student achievement would be useful in identifying segments of the population at greatest educational risk so that, once identified, actions could be taken to enhance their educational opportunities. Tyler (1966) argued forcefully that better information was needed to make wise decisions about policies and the allocation of resources, stating, for example, that

> the great educational tasks we now face require many more resources than have thus far been available, resources which should be wisely used to produce the maximum effect in extending educational opportunity and raising the level of education. To make these decisions, dependable information about the progress of education is essential . . . Yet we do not have the necessary comprehensive dependable data; instead, personal views, distorted reports, and journalistic impressions are the sources of public opinion. This situation will be corrected only by a careful, consistent effort to obtain data to provide sound evidence about the progress of American Education. (p. 95)

Equal educational opportunity was a major interest of Tyler's and of Francis Keppel in lending his support to the idea of a national assessment. Keppel, however, had in mind a more focused and precise instrument than Tyler or the developers ever thought possible. In testimony before the Select Committee on Equal Opportunity chaired by Senator Walter Mondale on December 1, 1971, Keppel gave enthusiastic support to NAEP and argued that the assessment movement fostered by NAEP had great potential utility for purposes of allocating resources to enhance both the quality and equality of educational opportunity.

> There is an extraordinary hopeful possibility that out of this movement we can develop measures by the school—the program within the school building—which will make it possible—not now, sir, but in due course—to rifle-shoot direct funds to improve the performance within a school building.

> I am making a contrast here between the school system as a whole—all the primary, junior high, and high schools, treated as a unit—because the important data on equal educational opportunity gets lost in the aggregate. It would seem to me essential that we disaggregate it; get the unit of measure down to the school itself, the place where the individual in charge can be held more responsible, in my judgment, than the superintendent. (Hearings before the Select Committee on Equal Educational Opportunity of the U.S. Senate, 1971, p. 10950)

The developers clearly had more modest expectations, especially after initial assessments made it clear how far removed the information was from Keppel's ambitious vision, and early evaluators faulted NAEP for its limitations. According to the staff response to the Greenbaum et al. (1977) evaluation, for example,

> census-like data, the planners knew even then, would not be very dramatic. People expecting quick and simple answers to fundamental questions (Why can't Johnny read?) would be disappointed with initial assessment results. (p. 199)

Womer and Mastie (1971) were even more circumspect:

> A recurring concern, both among those who support national assessment and those who have reservations about it, is the ultimate utility of the results. How will they affect education in this country? This is a very difficult question. While national assessment is designed to provide general information, it is not designed to produce answers to specific questions. (p. 118)

Demands that NAEP serve a wide variety of purposes, some of which it is ill equipped for (e.g., providing a basis for making strong causal inferences) and others of which may be in direct conflict for scarce resources (e.g., comparing current achievement to that of students in previous decades vs. assessing content that is considered most vital for the demands of the future), have created tensions throughout the history of NAEP. The nature of inferences that are appropriate to draw from NAEP results has been a continuing issue for NAEP that has been addressed repeatedly by evaluators. The final report of the National Academy of Education (NAE) evaluation, for example, summarized the issue as follows. "There is a natural inclination, particularly when there has been substantial investment . . . to relate NAEP achievement results to the instruction factors surveyed in NAEP's teacher and background questionnaires in order to draw inferences about what works and what fails to work in one's state, district or school. Unfortunately, NAEP is not suitable for drawing strong inferences about which factors or variables account for educational achievement" (Glaser, Linn, & Bohrnstedt, 1997, p. 19).

The Nature of the Assessment

Tyler made a strong case early on that the assessments should be developed using principles rather different than those used for standardized tests. He wanted the focus to be on individual items, referred to as exercises, not on a total score. This created a demand to develop exercises that would have high face validity. That is, educators and the general public should be able to see the performance on the exercises as attainments with clear educational and social value. Properties of items valued for a standardized test, such as an item's power to discriminate among individuals, were considered irrelevant for national assessment exercises.

Tyler also rejected the focus of standardized tests on grade-level norms. He argued that the order in which topics are taught was so variable across the nation that a grade-level comparison is not meaningful. In place of grade level Tyler argued that the assessments should be targeted to selected age levels, and initially proposed ages 9, 12, 15, and 18 years (Hazlett, 1974, pp. 27-28).

As has already been noted, Tyler was highly influential in shaping NAEP in the early years. The early assessment did indeed focus on age rather than grade cohorts and on individual exercises selected for their intrinsic interest rather than their ability to discriminate among individuals. Certainly there were refinements of the ideas and considerable labor involved in working out the details for the assessments between the time of Tyler's 1963 memorandum and the first administration of a NAEP by the Education Commission of the States (ECS) in 1969, but the effort was sent on a definite course by that memorandum. Many other individuals played major roles in fleshing out the broad outline through the work of the Exploratory Committee for Assessing the Progress of Education (ECAPE) between 1964 and 1968 and the reconstituted form of ECAPE as the Committee on Assessing the Progress of Education (CAPE) that functioned from October 1965 until the end of June 1969. The Technical Advisory Committee consisting of Robert Abelson, Lee Cronbach, Lyle Jones, and John Tukey (Chairperson), all of whom served from 1965 until 1969, was also critically important in developing plans for designing exercises, sampling, administration, and analysis, as well as in evaluating proposals and working with contractors (see, Fitzharris, 1993; Hazlett, 1974).

The ECS Years

The first NAEP administrations under the auspices of the Education Commission of the States (ECS) took place during 1969-1970. Three content

areas—science, writing, and citizenship—were assessed for students at ages 9, 13, and 17 years, for out-of-school 17 year-olds, and for adults 26-35 years old. The second round of NAEP administration for the same age groups took place in 1971 for the areas of reading and literature. That year was also the first that NAEP was financed solely by federal funding. Between 1972 and 1983, when the responsibility for the conduct of NAEP was moved to Educational Testing Service (ETS), assessments were conducted in mathematics, science, music, social studies, reading, art, writing, citizenship / social studies, basic life skills, and career and occupational development. During the ECS tenure, science was assessed four times, reading, mathematics, and social studies (either alone or in combination with citizenship) three times each, and music, art, and writing twice each.

The accomplishments of NAEP during the ECS years were impressive in many ways. The system produced trustworthy results that were able to document trends in achievement in several subjects. The awareness of NAEP by the public and by policymakers, however, was extremely limited. There were also questions and concerns expressed by those who were aware of NAEP results. “Trend patterns in some subject areas, like science, provided valuable and somewhat disturbing results . . . The [science] achievement results of seventeen year olds continued to decline across the three testing periods” (Fitzharris, 1993, p. 74). Some educators argued that the assessments failed to reflect current thinking in science due to the reuse of exercises prepared in the 1960s and administered in the mid to late 1970s as a means of measuring trends. Keeping NAEP aligned with a changing curriculum while measuring trends was more of a challenge than seems to have been anticipated. It is one that has continued to draw the attention of evaluators of NAEP starting in the late 1970s and continuing through the present.

Individual exercises. During the years that ECS was the prime contractor for NAEP, performance on individual assessment exercises / items was emphasized in reports. Hazlett (1974) summarized the rationale for this mode of reporting as follows. “It should be emphasized again that the National Assessment program emphasizes the importance of individual items as having intrinsic merit. An important aspect of National Assessment reports is therefore the release of individual items. In the first report of each subject matter area from 40% to 50% of all items used were released with their respective ‘p’ values” (p. 21).

Although the reporting of results for individual items is a straightforward way of reporting performance and has the advantage of directly showing what

proportion of students can do a particular concrete task, this style of reporting has limitations. The most obvious limitation is that it does not provide an overall summary of results. Averages of *p*-values, of course, can be computed for a fixed set of released items for the total sample or for designated subsamples. Averages for different collections of items are not comparable, however, and therefore trends in performance are only meaningful for individual items or for common sets of items administered in each of the assessments being compared. As additional items are released after each assessment, the subset of non-released items available for judging trends over several assessments is reduced to a smaller and smaller number of items with each successive assessment.

Fitzharris' (1993) account of the oversight of NAEP during the last half of the 1970 indicates that this was a period of transition from an atmosphere of a research-oriented assessment system to a regularized statistical survey in which there was increasing scrutiny of budgets and technical operations. "By the late 1970s, the administration of NAEP was no longer considered just the purview of the Education Commission of the States (ECS). It was clear the contract for 1983-1987 would be challenged" (Fitzharris, 1993, p. 81).

A report entitled *Measuring the Quality of Education: A Report on Assessing Education Progress*, prepared by Willard Wirtz and Archie Lapointe pursuant to grants from the Carnegie Corporation, the Ford Foundation, and the Spencer Foundation, was released in 1982 (Wirtz & Lapointe, 1982). The Wirtz and Lapointe report supported the continuation of NAEP, but made a number of recommendations that were subsequently adopted after a new contract for the conduct of NAEP was awarded to ETS. Among the recommendations subsequently followed as part of the ETS contract or later adopted by Congress were the recommendations that provisions be made for publishing state-by-state results, that results be reported on a grade-level basis for aggregations of exercises (rather than only individual exercises), and that an Assessment Policy Council be established to provide an overview function for NAEP (Wirtz & Lapointe, 1982).

The Educational Testing Service Proposal and Contract

In February 1983 it was announced that Educational Testing Service (ETS) had won the NAEP contract. The successful ETS proposal called for substantial changes in the design, analysis, and reporting of NAEP results. The broad outline for the new design and the rationale for the changes were articulated in a report entitled *National*

Assessment of Educational Progress Reconsidered: A New Design for a New Era (Messick, Beaton, & Lord, 1983).

Scale scores. One of the major changes that was introduced with the 1984 assessment conducted by ETS was the introduction of NAEP scale scores using item response theory (IRT). The NAEP scale scores were defined to have a theoretical range from 0 to 500, though typically falling between 100 and 400. The scale was defined as a developmental scale that spanned all three age levels assessed by NAEP and was interpreted as estimated scores on a hypothetical 500-item test with specified properties. The scale was defined originally to have a mean of 250 and a standard deviation of 50 for the combined sample from Grades 4, 8 and 12.

In comparison to the earlier reliance on *p*-values for common sets of items, the IRT scaling had the advantage that it provided a basis for tracking trends for assessments that had both common and unique sets of items. The common developmental scale across grade levels and ages also was thought to provide a means of comparing changes in achievement for a given age or grade level to the magnitude of the differences between age or grade levels. On the other hand, the scale was abstract and seemingly more difficult to interpret than *p*-values for individual items. The appropriateness of the cross-age scaling was challenged (see, for example, Forsyth, 1991; Haertel, 1989), and it was decided to discontinue cross-grade scaling in favor of within-grade and within-age scaling. Scale interpretation is a difficult challenge, however, whether the scaling is done within or between grade levels. Consequently, a variety of other approaches to aid in the interpretation of results have been attempted. These approaches rely, in part, on the use of exemplar items that are linked to the scale. Two of these approaches, item anchoring and the setting of achievement levels, are discussed below.

State-by-State NAEP

As previously indicated, some of the earliest conceptions for NAEP would have had results of the assessment reported at the state level. Because of concerns about local control and the concerns about federal review of state performance, however, decisions were made to limit the reporting of NAEP results to the nation as a whole, with additional reports for four broad regions and for subpopulations defined by demographic characteristics of the students assessed. Some states, such as Maine, did adopt the use of released NAEP items for their own state assessments, but

comparisons among states in terms of the performance of their students on NAEP were not possible.

From 1969 to 1983 NAEP was administered by ECS, an organization with a mission to serve states. It may be somewhat surprising then, particularly from a contemporary vantage point, that the NAEP continued to be confined to reports of national and regional results throughout that period. Not only do we now have the experience of almost a decade of successful state-by-state administration of NAEP and reporting of NAEP results; President Bush's education proposal includes plans to use annual state-level NAEP results for purposes of state-level accountability (The White House, 2001). During this ECS period, however, the concerns about state comparisons remained sufficient to make such a policy consistent with the desires of the states that ECS serves.

As was noted by Fitzharris (1993), the concerns about federal control and state comparisons had become less of an issue by the early 1980s. The reduction in concern was foreshadowed in the Congressional Act of 1978 (Public Law 95-561) that reauthorized NAEP. "The legislation stated that a responsibility of the National Assessment was to 'provide technical assistance to State educational agencies and to local educational agencies on the use of National Assessment objectives' " (Fitzharris, 1993, p. 110). The movement toward actual state comparisons based directly on NAEP was accelerated by the publication in 1983 of *A Nation at Risk* (National Commission on Excellence in Education, 1983), which concluded, based in part on reference to data from NAEP, that poor academic achievement placed the nation "at risk." Publication of "Wall Charts" of statistics comparing states in terms of average scores on the Scholastic Aptitude Test (SAT) and the American College Test (ACT), along with other statistics by Secretary of Education Terrel Bell, further accelerated the move toward the reporting of state-by-state NAEP results.

Comparisons of states based on SAT or ACT scores have many flaws, among the most obvious of which is that the proportion of students taking one of the tests varies greatly from state to state (see, for example, Linn, 1987). Moreover, the students who take those tests are, for the most part, only students who plan to attend college and thus are not representative of all students at an age or grade level. If state comparisons based on student test performance were to be made, it was clear that a more valid basis of comparison was needed. NAEP was turned to as providing the best basis for making such comparisons. In 1984 the Council of Chief State School Officers (CCSSO) recommended that the law authorizing NAEP be

changed to permit the use of NAEP to make state-by-state comparisons (Fitzharris, 1993).

In May 1986, Secretary of Education William Bennett formed a blue-ribbon study group chaired by Lamar Alexander, the governor of Tennessee, to review NAEP and to recommend possible changes in its roles and conduct. J. Thomas James, president emeritus of the Spencer Foundation, was appointed as the study director. The study group submitted its report to Secretary Bennett in January 1987 (Alexander & James, 1987). That report, which is commonly referred to as the Alexander-James report, contained a number of far-reaching recommendations, several of which had a direct influence on the legislation that reauthorized NAEP in 1988 (Public Law 100-297).

Given the context of the Wall Charts and the recommendation of the CCSSO, it is not surprising that the Alexander-James report gave a high priority to changing NAEP to allow state-by-state comparisons. Indeed, the study group made state-by-state comparisons its number one recommendation. “The single most important change recommended by the Study Group is that the assessment collect representative data on achievement in each of the fifty states and the District of Columbia. Today, state and local school administrators are encountering a rising public demand for thorough information on the quality of their schools, allowing comparison with data from other states and districts and with their own historical records” (Alexander & James, 1987, pp. 11-12).

The study group acknowledged that there had been concerns in the past about comparing states in terms of achievement results produced by an assessment such as NAEP, but concluded that those “concerns are less important now than they were previously, and that most can be readily accommodated with a redesigned national assessment” (Alexander & James, 1987, p. 5). A major element of the study group’s rationale for state-level reporting of NAEP results was that states and local districts have the responsibility for making the most important decisions in education and they need information that they lacked to see how well their schools were doing.

Some concerns were expressed by the National Academy of Education (NAE) panel that was formed to provide commentary on the Alexander-James report. The NAE panel, which was chaired by Robert Glaser, expressed concern about emphasis that might be given to ranking states on the basis of their scores on NAEP. “We are concerned about the emphasis in the Alexander-James report on state-by-state

comparisons of average test scores. Many factors influence the relative ranking of states, districts, and schools. Simple comparisons are ripe for abuse and are unlikely to inform meaningful school improvement efforts" (Glaser, 1987, p. 59). In general, however, there seemed to be fairly widespread support for the recommendation that NAEP be used to obtain state-by-state results.

Congress followed the advice of the Alexander-James study group when it enacted the reauthorization of NAEP in 1988 (Public Law 100-297). The reauthorization included an option for states to obtain state-level NAEP results by participating in the Trial State Assessment (TSA) as part of the 1990 and 1992 administrations of NAEP. The 1988 reauthorization also included a requirement for the evaluation of the TSA by either the NAE or the National Academy of Sciences (NAS).

The 1990 TSA was limited to the eighth-grade mathematics assessment. State participation was voluntary. Thirty-seven states, the District of Columbia, and two territories participated in the assessment. Unlike the national samples, which included students in both public and private schools, the state samples were limited to public schools. Responsibility for the evaluation of the TSA was given to the NAE. The NAE formed a panel to conduct the evaluations of the 1990 and the 1992 TSAs. The first NAE panel report focusing on the evaluation of the 1990 TSA was issued in 1992 (Glaser, Linn, & Bohrnstedt, 1992). The report concluded that the 1990 TSA had been carried out successfully. Since the 1990 TSA had been limited to one grade and one subject, and because the 1992 TSA was already underway at the time the report was released, the report recommended that the trial be continued in an expanded form as part of the 1994 assessment.

The NAE panel also made a number of recommendations for improving state NAEP. Four of the panel's recommendations were adopted. These were:

1. the trial should be continued for one more round in 1994 (adopted but due to funding limitations on a smaller scale than the panel had recommended);

2. the state samples should include students attending private schools (adopted by Congress);

3. the practice of random monitoring of schools participating in state NAEP should be continued (adopted); and

4. new reporting devices should be used and the release of results should be spread out throughout the year (adopted).

Two other recommendations were followed, at least in part. These were the recommendation that NAEP content be comprehensive and reflect the most up-to-date approaches as well as current classroom instruction, and that the exclusion of children with disabilities be evaluated. The remaining five recommendations of the panel were not adopted. These included the recommendation that out-of-school 17 year olds be included in the assessment (that had been the case in the earliest years of NAEP), and that the prohibition against reporting results below the state level be continued. The prohibition was, in fact, lifted and districts as well as states are now allowed to participate in NAEP.

The second administration and reporting of NAEP at the state level was in 1992. In 1992, the Trial State Assessment (TSA) again assessed mathematics at the eighth grade. It also assessed reading and mathematics for the first time at the fourth grade. The number of states participating increased from 37 in 1990 to 41 in 1992. The evaluation of the 1992 TSA by the NAE panel followed up on many of the issues studied initially in its evaluation of the 1990 TSA. The bottom-line recommendation of the NAE panel, based on its evaluation of the 1990 and 1992 trials, was "that the Congress authorize a continuation of state NAEP" (Glaser, Linn, & Bohrnstedt, 1993, p. 104). The panel also recommended that "Congress mandate ongoing evaluation of state NAEP with ongoing feedback to Congress" (p. 104).

Congress extended the TSA to include the 1994 administration of NAEP and the NAE panel's evaluation was expanded to include the 1994 TSA of fourth-grade students in reading (Glaser, Linn, & Bohrnstedt, 1996). Subsequently, Congress authorized the continuation of state NAEP and did mandate continued evaluation, mandating an evaluation of the national assessment as well as the state assessment (PL 103-382). That mandated evaluation was conducted by a committee of the National Research Council (NRC; Pellegrino, Jones, & Mitchell, 1999).

The NAE panel evaluations of the 1992 and 1994 trials devoted considerable attention to issues of the content and cognitive demands of the assessments, the issue of inclusion of students with disabilities and students with limited English proficiency in the assessments, and ways to make sampling for the state assessments more efficient. A number of recommendations were made for ways to ensure that the content of the assessments would be comprehensive, "reflecting the most forward-looking pedagogical approaches, at the same time [as] they reflect the best of current practice" (Glaser et al., 1993, p. xxii). Arguably, NAGB, NCES, and the

NAEP contractor have worked to achieve that recommendation to the extent feasible within the constraints of the resources that are available for NAEP.

Inclusion of students in assessments is a major issue not only for NAEP but for state- or district-mandated assessments, as well. Inclusion of students with disabilities in assessments is a requirement of the 1997 amendments to the Individuals with Disabilities Education Act (IDEA) of 1975 (PL 94-142). The IDEA requirement to provide students with accommodations so they can meaningfully participate in assessments is clear. Both the desirability of inclusion and the substantial unknowns of including students in meaningful ways that produce valid results were recognized by the NAE panel in its recommendations for NAEP. "NCES and NAGB should continue efforts to encourage greater participation of student with disabilities or limited English proficiency in the current NAEP. At the same time, they should continue research to identify adaptations or accommodations for each of these groups that would provide more valid measures of subject-area achievement as specified by the NAEP frameworks" (Glaser et al., 1996, p. xxiii). The National Center for Education Statistics (NCES) and the National Assessment Governing Board (NAGB) have made major efforts both to make NAEP more inclusive and to carry out the data collection and analyses needed to preserve the validity and interpretability of results, particularly trends in achievement.

The Formation of NAGB

Throughout the period that NAEP was conducted by the Education Commission of the States (ECS) and for the initial years of the ETS contract, the governance of NAEP was under the direct control of the Office/Department of Education. With input from the contractor, the Department appointed a council to address policy issues. The contractor coordinated the work of the council. Technical advisory bodies were appointed by and provided advice directly to the contractor.

The Alexander-James panel had reservations about the lack of a governance structure that was independent of the National Center for Education Statistics (NCES), the Department of Education, and the contractor. "In order to undertake such demanding new tasks as state-by-state comparisons, the national assessment will require some important changes in its current governance structure. . . . We recommend the creation of an independent governing agency, the Educational Assessment Council (EAC). The EAC would operate independently of the institution carrying out the actual assessment and would define content areas,

assessment procedures, and guidelines for fair comparisons of states and localities" (Alexander & James, 1987, p. 13).

The 1988 reauthorization of NAEP (PL 100-297) established a new governance structure for NAEP. NCES retained responsibility for operations and technical aspects of NAEP analyses. Responsibility for governing NAEP was assigned by a newly authorized body, the National Assessment Governing Board (NAGB). Among its responsibilities, NAGB selects the subject areas to be assessed and is responsible for the NAEP frameworks that specify the content of the assessments in each subject area.

Reporting NAEP Results

The way that NAEP results are reported has been the focus of considerable attention both by those responsible for NAEP and by those who have evaluated it. We have already noted that the reporting of results in terms of individual exercises was considered inadequate for the needs of the assessment in the ETS proposal. The scale scores avoided some of the limitations of only reporting the proportion of students who responded correctly to individual items or average *p*-values for defined subsets of items, but the meaning of the scale was obscure. Consequently, efforts were made to make the results more interpretable. One way of doing this was by identifying items that became known as item anchors associated with selected locations on the scale. The intent was to give meaning to the scale by showing items located at selected points on the scale that students would be likely to answer correctly while lower scoring students could not. Another approach to giving more meaning to the NAEP scale has been the identification of achievement levels on the scale that correspond to levels of achievement that are judged to be what students should be able to do.

Item anchoring. The first report of NAEP results after ETS became the prime contractor was in 1984. To aid in interpretation, ETS associated labels with scale scores at 50-point intervals from 150 to 350, ranging from rudimentary (150) to advanced (350). "Users did not find this labeling helpful, because of the cross-grade scale" (Shepard, Glaser, Linn, & Bohrnstedt, 1993, p. 18). Such labeling was also considered to be beyond the purview of the contractor. Hence the practice was discontinued. Instead, item anchoring was used in an effort to give meaning to the scale.

Item anchoring procedures were developed by Albert Beaton while he was the head of the data analysis team at ETS. Items were selected as anchors for the scale points of 150, 200, 250, 300, and 350 based on their statistical properties. In order to qualify as an anchor item, an item had to be answered correctly by at least 65% of the students whose overall performance placed them at the anchor point. In addition, the item had to be answered correctly by less than half the students at the next lower anchor point, and the difference between the percentage of students located at the target anchor point must be at least 30 points higher than the corresponding percentage at the next lower anchor point. Once the eligible items were selected for each anchor point using these statistical criteria, they were carefully examined by a group of content specialists and used as the basis for writing substantive anchor-point descriptions of what students scoring at that point or above were able to do on the assessment. Exemplar anchor items were also selected for use as explicit illustrations of items answered correctly by two thirds or more of the students at the anchor point, but less than half the students at the next lower anchor point.

When item anchors were shown as exemplars, they were accompanied by a report of the percentage of students who scored above that point. Unfortunately, as noted by Forsyth (1991) and by Linn and Dunbar (1992), popular interpretations in the media erroneously assumed that because, for example, 19% of the 13-year-old students scored above an anchor score of 300, that only 19% could correctly answer the anchor item reported to exemplify performance at a score of 300. The actual percentage of students who answered the item correctly in this example was 51%. As explained by Linn and Dunbar (1992) the much larger percentage actually answering the anchor item correctly was due to the fact that substantial numbers of students scoring below the anchor point on the scale as well as those scoring above that point answered the item correctly. Issues about the use of item anchors became largely moot when NAGB decided to set performance standards, called achievement levels, and use them in the reporting of NAEP results.

Achievement levels. The idea for reporting NAEP results in terms of a small number of descriptive categories was suggested by the NAE panel that was formed to provide commentary on the Alexander-James report. The NAE panel recommended “that to the maximal extent technically feasible, NAEP use descriptive classifications as the principal scheme in future assessment. For each content area NAEP should articulate clear descriptions of performance levels,

descriptions that might be analogous to such craft rankings as novice, journeyman, highly competent and expert" (Glaser, 1987, p. 58). This idea of descriptive categories was quite consistent with notions of criterion-referenced measurement as conceptualized a quarter of a century earlier by Glaser (1963).

The NAGB first started its effort to report results in terms of performance standards with the 1990 mathematics assessment. The development and use of performance standards to report NAEP results was undertaken early in the tenure of the then newly established Governing Board. The performance standards that were established by NAGB were named "achievement levels."

The decision to report NAEP results in terms of achievement levels was based on the Governing Board's interpretation of the legislation that reauthorized NAEP. Among other responsibilities, the legislation assigned NAGB the responsibility of "identifying appropriate achievement goals" (PL 100-297, Part C, Section 3403 (6) (A), 1988). As was noted by the National Academy of Education (NAE) Panel on the Evaluation of the NAEP Trial State Assessment (Shepard et al., 1993), the Board "might have responded in different ways. Given the emerging consensus for establishing national education standards, the fact that the Education Summit was silent on who should set standards, and the fact that NAEP was the only national assessment of achievement based on defensible samples, NAGB interpreted the authorizing legislation as a mandate to set performance standards, which it named 'achievement levels,' for NAEP" (p. xviii).

Because of the potential importance of the achievement levels in the context of the press for national standards, the 1990 achievement levels were subjected to several evaluations (e.g., Linn, Koretz, Baker, & Burstein, 1991; Stufflebeam, Jaeger, & Scriven, 1991; U.S. General Accounting Office, 1992, 1993). NAGB was responsive to many of the criticisms of the evaluators and undertook another, more extensive standard-setting effort for the 1992 mathematics and reading assessments. Evaluations of the 1992 effort (e.g., Burstein et al., 1993; 1995/1996; Shepard, 1995; Shepard et al., 1993), however, were again quite critical. The NAE panel, for example, concluded that the achievement levels might reduce rather than enhance the validity of interpretations of NAEP results. Not all agreed. Indeed, there were strong defenders of the 1992 mathematics achievement levels, the process used to set them, and the interpretations they were intended to support (e.g., American College Testing, 1993; Cizek, 1993; Kane, 1993).

The controversy, at least in part, led to a conference on standard setting for large-scale assessments that was held in October 1994 under the joint sponsorship of NAGB and NCES. Although the conference did not resolve the controversy, several of the papers, which are available in the conference proceedings (NAGB & NCES, 1995), clarified aspects of the controversy. Much of the debate about achievement levels has focused on the standard setting method. That is hardly surprising given that the NAE panel recommended against the use of "the Angoff method or any other item-judgment method to set achievement levels" because the panel concluded that such methods are "fundamentally flawed" (Shepard et al., 1993, p. 132).

There is broad agreement that different methods of setting standards lead to different cut-scores, but there is not general agreement that one particular method is best. Given this impasse in the technical community, together with agreement that standards ultimately involve policy judgments, NAGB has continued to rely on achievement levels using techniques considered flawed by some evaluators.

The most recent evaluations of NAEP, conducted by a committee of the National Research Council (NRC), have continued to find fault with the NAEP achievement levels. This is evident in summary recommendation 5 of the NRC committee on the evaluation of NAEP: "Summary Recommendation 5. The current process for setting achievement levels should be replaced. New models for setting achievement levels should be developed in which the judgmental process and data are made clearer to NAEP users" (Pellegrino et al., 1999, p. 162). This NRC recommendation, like similar ones before it from other evaluators, was rejected by NAGB.

The analysis of the NAEP achievement levels is one of the few areas where the conclusions of evaluators of NAEP have been uniformly and consistently negative without having any impact on practice. In this area, recommendations of evaluators have been ignored or rejected. An unfortunate side effect of this impasse is that it has tended to overshadow other aspects of evaluation reports so that some of the other conclusions and recommendations of evaluation panels may have had less impact than they otherwise might have had. This appears to be the case, for example, for the report of the NRC Committee on the Evaluation of the National and State Assessments of Educational Progress (Pellegrino et al., 1999). The NRC committee reached five broad summary conclusions and made broad recommendations for each of those conclusions, only one of which dealt with

achievement levels, but it was the conclusion and associated recommendation on achievement levels that received the most attention when the report was presented. The other four summary conclusions and associated recommendations are potentially important ones for the future of NAEP.

Coordinated System of Indicators

The first summary conclusion concerned the limited nature of any single data collection such as NAEP in measuring all the aspects of student achievement and the varied needs for information about education. Associated with this conclusion was the NRC committee's first summary recommendation.

> The nation's educational progress should be portrayed by a broad array of education indicators that includes but goes beyond NAEP's achievement results. The U.S. Department of Education should integrate and supplement the current collections of data about education inputs, practices, and outcomes to provide a more comprehensive picture of education in America. In this system, the measurement of student achievement should be reconfigured so that large-scale surveys are but one of several methods used to collect information about student achievement. (Pellegrino et al., 1999, p. 22)

The earlier report of the NAE panel had made suggestions that there needed to be better integration with other sources of information and that NAEP should serve as a resource for other data collection systems (Glaser et al., 1997, pp. 108-109). The NRC committee developed the idea of a coordinated data collection system more fully, however. The coordinated system envisioned by the NRC committee would include information on financial resources, school organization and governance, teacher and professional development, instructional practice, content standards and curricula, school climate environment, home and community support for learning, and student background, as well as student achievement. In addition to NAEP, with its state and national assessments, the student achievement component of the coordinated system would also include international assessments such as the Third International Mathematics and Science Study, and student-level longitudinal data collections such as the National Education Longitudinal Study and the Early Childhood Longitudinal Study.

Redesign

In its second summary recommendation, the NRC committee proposed a significant reconfiguration of NAEP.

> Summary Recommendation 2. NAEP should reduce the number of independent large-scale data collections while maintaining trend lines, periodically updating frameworks, and providing accurate national and state-level estimates of academic achievement. (Pellegrino et al., 1999, p. 56)

In the view of the committee the reduction of independent large-scale data collections could be accomplished, in part, by merging the main NAEP with the trend NAEP. As noted in the committee report, this recommendation is consistent with suggestions for redesign offered by NAGB (1997), the NAGB Design Feasibility Team (Forsyth, Hambleton, Linn, Mislevy, & Yen, 1996), and the NAE panel's final report (Glaser et al., 1997), and some steps toward that goal have been taken.

The NRC committee also argued that the national and state designs could be streamlined by combining them to the extent feasible. This suggestion has also been made by other groups, but operational obstacles have prevented any clear action in this direction.

In addition to recommending streamlining and combining the large-scale survey functions of NAEP, the NRC committee concluded that the "collection of meaningful NAEP data in the twelfth grade is problematic given the insufficient motivation of high school seniors and their highly variable curricula and dropout rates" (Pellegrino et al., 1999, p. 84). Therefore, the committee recommended that alternatives of assessing students in Grades 10 or 11 be explored.

Inclusion

Consistent with urgings of the NAE panel on the evaluation of the TSA, the NRC committee's third summary conclusion focused on the assessment of students with disabilities and English language learners.

> Summary Conclusion 3. NAEP has the goal of reporting results that reflect the achievement of all students in the nation. However, many students with disabilities and English-language learners have been excluded from the assessments. Some steps have been taken recently to expand the participation of these students in NAEP, but their performance remains largely invisible. (Pellegrino et al., 1999, p. 87)

The conclusion and the associated recommendation that NAEP should continue to strive for the meaningful inclusion of students with disabilities, like similar recommendations from others earlier, has consistently been responded to with good-faith efforts on the part of NCES and NAGB. The reality is, however, that the field of educational measurement has not gotten very far with approaches that

provide valid ways of including students with some types of disabilities or English language learners in assessments so that their results can be combined with, or meaningfully be compared with, results of other students. The challenges in this area of assessment are many, and the solid answers are relatively few. Hence, the meaningful inclusion of all students in assessments, while required by law and recommended by many different groups evaluating assessments, remains a substantial challenge, not just for NCES and NAGB, but for the field of educational assessment generally.

One unique aspect of the NRC committee recommendation was linked to the committee's recommendation for a part of a coordinated system of educational indicators. Specifically, the committee recommended that "the proposed system of education indicators should include measures that improve understanding of the performance and educational needs of these populations" (Pellegrino et al., 1999, p. 87).

Content

The fourth summary conclusion and associated recommendation made by the NRC evaluation committee harkened back to a theme found repeatedly in evaluations, namely that the content coverage of NAEP should not only be comprehensive, but should reflect what is known about learning and achievement. This theme is consistent with that found in the earlier NAE panel reports and in the Technical Review Panel report (Haertel, 1991). Specifically, the NRC committee's fourth summary recommendation was:

> The entire assessment development process should be guided by a coherent vision of student learning and by the kinds of inferences and conclusions about student performance that are desired in reports of NAEP results. In this assessment development process, multiple conditions need to be met: (a) NAEP frameworks and assessments should reflect subject-matter knowledge; research, theory, and practice regarding what students should understand and how they learn; and more comprehensive goals for schooling; (b) assessment instruments and scoring criteria should be designed to capture important differences in the levels and types of students' knowledge and understanding both through large-scale surveys and multiple alternative assessment methods; and (c) NAEP reports should provide descriptions of student performance that enhance the interpretation and usefulness of summary scores. (Pellegrino et al., 1999, p. 114)

The NRC committee's vision of the content of NAEP is ambitious. It is one that NCES and NAGB arguably try to achieve. Compromises due to considerations of

assessment time, cost, and the need to maintain trends, however, inevitably make the assessment fall short of the vision.

Conclusion

NAEP has been subjected to close scrutiny by many evaluation panels and by individual investigators throughout its history. There have been numerous recommendations from the many evaluations. The preponderance of those recommendations have been heeded to the extent feasible within the constraints of programs imposed by the level of authorizations. A few recommendations have been rejected because they were considered inconsistent with desired policy or because they were deemed infeasible.

Although evaluators have not hesitated to be critical of many aspects of NAEP, the general tenor of the evaluators' conclusions has been quite positive. NAEP has been subjected to close scrutiny and found to be a valuable source of information about trends in student achievement for states and the nation.

References

Alexander, L., & James, T. (1987). *The nation's report card: Improving the assessment of student achievement. Report of the study group.* Cambridge, MA: National Academy of Education.

American College Testing Program. (1993). *Setting achievement levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading and Writing: A technical report on reliability and validity.* Iowa City, IA: American College Testing Program.

Burstein, L., Koretz, D. M., Linn, R. L., Baker, E. L., Sugrue, B., Novak, J., et al. (1995/1996). Describing performance standards: The validity of the 1992 NAEP achievement level descriptors as characterizations of mathematics performance. *Educational Assessment, 3,* 9-51.

Burstein, L., Koretz, D. M., Linn, R. L., Sugrue, B., Novak, J., Lewis, E., et al. (1993). *The validity of interpretations of the 1992 NAEP achievement levels in mathematics* (Technical Report). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cizek, G. J. (1993). *Reactions to National Academy of Education report: Setting performance standards for student achievement.* Washington, DC: National Assessment Governing Board.

Fitzharris, L. H. (1993). *An historical review of the National Assessment of Educational Progress from 1963 to 1991.* Unpublished doctoral dissertation, University of South Carolina.

Forsyth, R. A. (1991). Do NAEP results yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice, 10*(3), 3-9, 16.

Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., & Yen, W. (1996). *Design/feasibility team. Report to the National Assessment Governing Board.* Washington, DC: National Assessment Governing Board.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-521.

Glaser, R. (1987). *Commentary by the Nation's Academy of Education, Part 2 of the Nation's Report Card: Improving the assessment of student achievement. Report of the study group.* Cambridge, MA: National Academy of Education.

Glaser, R., Linn, R., & Bohrnstedt, G. (1992). *Assessing achievement in the states.* Stanford, CA: The National Academy of Education.

Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *The trial state assessment: Prospects and realities.* Stanford, CA: The National Academy of Education.

Glaser, R., Linn, R., & Bohrnstedt, G. (1996). *Quality and utility: The 1994 trial state assessment in reading.* Stanford, CA: The National Academy of Education.

Glaser, R., Linn, R., & Bohrnstedt, G. (1997). *Assessment in transition: Monitoring the nation's educational progress.* Stanford, CA: The National Academy of Education.

Greenbaum, W., Garet, M. S., & Solomon, E. (1975). *Measuring educational progress.* New York: McGraw-Hill.

Haertel, E. (Chair). (1989). *Report of the NAEP technical review panel on the 1986 reading anomaly, the accuracy of NAEP trends, and issues raised by state-level NAEP comparisons* (NCES Tech. Rep. CS 89-499). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Hazlett, J. A. (1974). *A history of the National Assessment of Educational Progress, 1963-1973: A look at some conflicting ideas and issues in contemporary American education.* Unpublished doctoral dissertation. University Microfilms International No. AAC 7506135.

Kane, M. (1993). *Comments on the NAE evaluation of the NAGB achievement levels.* Washington, DC: National Assessment Governing Board.

Linn, R. L. (1987). Accountability: The comparison of educational systems and the quality of test results. *Educational Policy, 1,* 181-198.

Linn, R. L. (1988). State-by-state comparisons of achievement: Suggestions for enhancing validity. *Educational Researcher, 17*(3), 6-9.

Linn, R. L., & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29,* 177-194.

Linn, R. L., Koretz, D., & Baker, E. L. (1995). *Assessing the validity of the National Assessment of Educational Progress: Final report of the NAEP Technical Review Panel.* Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. L., Koretz, D., Baker, E. L., & Burstein, L. (1991). *The validity and credibility of the achievement levels for the 1990 National Assessment of Educational Progress in Mathematics.* Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Messick, S., Beaton, A., & Lord, F. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era* (Report 83-10). Princeton, NJ: Educational Testing Service.

National Assessment Governing Board. (1997). *Bridging policy to implementation: A resolution.* Washington, DC: National Assessment Governing Board.

National Assessment Governing Board and National Center for Education Statistics. (1995). *Proceedings of the Joint Conference on Standard Setting for Large-Scale*

Assessments of the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES) (Volume II). Washington, DC: Author.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. A report to the nation and the Secretary of Education. Washington, DC: U.S. Government Printing Office.

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.

Shepard, L. A. (1995). Implications for standard setting of the National Academy of Education evaluation of the National Assessment of Educational Progress achievement levels. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES)* (Volume II, pp. 143-160). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement.* Stanford, CA: The National Academy of Education.

Stufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1991). *Summative evaluation of the National Assessment Governing Board's inaugural effort to set achievement levels on the National Assessment of Educational Progress.* Kalamazoo: Western Michigan University, The Evaluation Center.

Tyler, R. W. (1966). The development of instruments for assessing educational progress. In *Proceedings of the 1965 invitational conference on testing problems* (pp. 95-105). Princeton, NJ: Educational Testing Service.

U.S. General Accounting Office. (1992). *National assessment technical quality* (GAO / PEMD-92-22R). Washington, DC: Author.

U.S. General Accounting Office. (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations* (GAO / PEMD-93-12). Washington, DC: Author.

The White House. (2001). *Transforming the federal role in education so that no child is left behind.* Washington, DC: The White House.

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

Womer, F. B. (1970). *What is national assessment?* Ann Arbor, MI: National Assessment of Educational Progress.

Womer, F. B., & Mastie, M. M. (1971). How will National Assessment change American education? An assessment of assessment by the first NAEP Director. *Phi Delta Kappan*, *53,* 118-120.