Assessing the Validity of the National Assessment of Educational Progress: NAEP Technical Review Panel White Paper

CSE Technical Report 416

Robert L. Linn, CRESST/University of Colorado at Boulder
Daniel Koretz, CRESST/RAND
Eva L. Baker, CRESST/University of California, Los Angeles

U.S. Department of Education
National Center for Education Statistics
Grant RS90159001

June 1996

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 90024-1522
(310) 206-1532

The work reported herein was supported in part under the National Center for Education Statistics Contract No. RS90159001 as administered by the U.S. Department of Education, Office of Educational Research and Improvement.

ASSESSING THE VALIDITY OF THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS: THE NAEP TECHNICAL REVIEW PANEL WHITE PAPER

Robert L. Linn, CRESST/University of Colorado at Boulder
Daniel Koretz, CRESST/RAND
Eva L. Baker, CRESST/University of California, Los Angeles

Abstract

During the past six years, under a contract from the National Center for Education Statistics, a Technical Review Panel has overseen and conducted a series of research studies addressing a range of validity questions relevant to various uses and interpretations of the National Assessment of Education Progress (NAEP). Study topics included the quality of NAEP data, the number and character of NAEP scales, the robustness of NAEP trend lines, the trustworthiness of and interpretation of group comparisons, the validity of interpretations of NAEP anchor points and achievement levels, the linking of other test results to NAEP, the effects of student motivation on performance, the adequacy of NAEP data on student background and instructional experiences, and what is understood from NAEP reports by educators and policy makers. This report describes the questions addressed by each study and summarizes the most important findings. In addition, general conclusions based on this body of research are presented and related to the major purposes of NAEP. A general conclusion is that the evolving and growing range of uses to which NAEP is put will create the need for ongoing validation work of the sort illustrated by the Panel's studies.

Purpose

The National Assessment of Educational Progress (NAEP) is a Congressionally-mandated project of the National Center for Education Statistics (NCES) that has provided periodic measures of the achievement of the nation's students for the past quarter century. The formulation of policy guidelines for NAEP is the responsibility of the National Assessment Governing Board (NAGB). NAGB was created in the 1988 reauthorization of NAEP (P.L. 100-297) and charged with a number of responsibilities including (as specified in the 1994 reauthorization, P.L. 103-382) the selection of subject areas to be assessed and

the development of appropriate student performance levels, assessment objectives, test specifications, guidelines for reporting, and standards for interstate, regional and national comparisons.

The Commissioner of Education Statistics is responsible for carrying out NAEP, with the advice of NAGB and the Advisory Council on Education Statistics through competitive awards to contractors. NAEP is intended to be conducted so that it provides "a fair and accurate presentation of educational achievement in reading, writing and other subjects included in the third National Education Goal, regarding student achievement and citizenship" (P.L. 103-382, Sec. 411 [b][1]).

In the 1988 reauthorization of NAEP the Commissioner was also given responsibility for providing "continuing reviews of the national Assessment, including validation studies . . ." (P.L. 100-297, Sec. 3403 [i][9][A]). As part of its effort to fulfill the Commissioner's responsibility for continuing review and validation studies, NCES published a Request for Proposals (RFP) to create a Technical Review Panel (TRP) that would provide technical reviews and conduct a series of validation studies. A contract was awarded to the Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA in conjunction with the University of Colorado at Boulder and RAND in 1989 to carry out that work.

During the past six years the TRP has conducted a series of studies on specific questions relating to the validity of the interpretations of NAEP results, including the quality of NAEP data, the number and character of NAEP scales, the robustness of NAEP trend lines, the trustworthiness of and interpretation of group comparisons, the validity of interpretations of NAEP anchor points and achievement levels, the linking of other test results to NAEP, the effects of student motivation on performance, the adequacy of NAEP data on student background and instructional experiences, and what is understood from NAEP reports by educators and policy makers. The TRP has also provided NCES with reviews of draft NAEP reports and technical advice on design, administration, analysis, and reporting issues.

The purposes of this report are (a) to provide a synthesis of what has been learned from the various studies of disparate technical issues and experience in reviewing NAEP materials, procedures, and reports and (b) to provide general

conclusions and recommendations based on that experience related to NAEP design, administration, analysis, and reporting issues.

Background of the TRP

NAEP has certainly had its full share of evaluations (e.g., Alexander & James, 1987; Greenbaum, Garet, & Solomon, 1977; National Academy of Education, 1992, 1993b), and the NAEP contractors and others have conducted a host of studies bearing on the validity of uses and interpretations of the results during its 25-year history. It was neither a shortage of external evaluations nor a lack of studies bearing on the validity of NAEP that led to the formation of the TRP. Rather, it was experience with problems encountered with the 1986 reading assessment and the positive reactions to the use of an ad hoc advisory group called the "NAEP Technical Review Panel on the 1986 Reading Anomaly, the Accuracy of NAEP Trends, and Issues Raised by State-Level NAEP Comparisons" (Haertel, 1989) that provided a model for the creation of a continuing TRP.

The initial analysis of the 1986 reading assessment data by the NAEP contractor, Educational Testing Service (ETS), revealed a large, abrupt, and pervasive decline in performance of 17-year-olds, and a smaller (but still quite large compared to changes encountered between assessments conducted between 1971 and 1984) decline for 9-year-olds from 1984 to 1986. The initial analyses for 13-year-olds, on the other hand, revealed a slight increase in reading proficiency. This pattern of results was considered to be anomalous by ETS staff and their Design and Analysis Committee (DAC). Both ETS and the DAC recommended that reporting should be delayed until analyses could be completed so that possible artifacts could be evaluated and the results could be better understood. NCES concurred with this recommendation and a major series of analyses were undertaken by ETS (see Beaton, Zwick, and collaborators, 1990) that eventually led to revised estimates of 1986 reading proficiencies at all three age levels and to changes in the assessment design for 1988 and for tracking long-term trends in future assessments.

Because of the central importance of accurate trends to the purposes of NAEP, the 1986 reading anomaly was treated with the utmost seriousness by ETS and NCES. It provided the occasion for careful analysis and rethinking of a number of design and analysis issues. NCES sought review and analysis external

to the Center or its contractors. It is in this context that NCES formed the Technical Review Panel on the NAEP Reading Anomaly in December 1987 with Edward Haertel as chairperson. The Haertel Panel was charged with examining the "apparent lack of comparability between the findings of the 1984 and 1986 reading assessments" and the "accuracy of NAEP trend data" (Haertel, 1989, p. iii). The Panel was also asked to address issues likely to arise in the expansion of NAEP to include state-level reporting of NAEP results.

The Haertel Panel concluded that the bulk of the apparent declines in the initial estimates of 9- and 17-year-old students' reading proficiencies were "probably artifactual." Although a number of criticisms and suggestions were provided by the Panel, it "generally endorsed the ETS investigation of the anomaly" (p. ix). The Panel concluded that NAEP provides the best available indicator of national trends in student achievement. The Panel also concluded, however, that the quality of NAEP's measurement of trends could be substantially improved and provided a number of recommendations toward that end.

It is notable that many of the factors that were cited as possible contributors to the anomaly—for example, subtle changes in administration procedures or context effects due to different placement of items within blocks from one assessment to the next—could have influenced many of the NAEP assessments. What set the anomaly apart was its magnitude and the fact that the results were strikingly inconsistent with expectations and other data. A major lesson of the anomaly is not merely that large problems may occasionally arise, but also that a host of factors may at any time influence the results of the NAEP and the validity of inferences drawn from it.

Occurring as it did so near the time of the 1988 reauthorization of NAEP and the legislative mandate for the Commissioner of Education Statistics to obtain reviews and validation studies, the experience with the 1986 reading anomaly and the Haertel Panel led naturally to the creation of a continuing Technical Review Panel. That experience also helped shape the priorities in the RFP when it was issued in the spring of 1989.

The CRESST proposal framed the proposed work of the TRP in terms of a broad conception of validity reflected in the thinking of major theorists such as Cronbach (1988, 1989) and Messick (1989). This framework makes it clear that

validity depends not only on the tasks used in an assessment or on the ways in which samples are drawn and assessments are administered, but on the uses and interpretations that are made of assessment results. An assessment might have a high degree of validity for one interpretation (e.g., the mathematics achievement of 9-year-old students improved from 1986 to 1990) yet be quite invalid for another interpretation (e.g., the higher achievement of students in state A than in state B is due to more stringent teacher certification requirements in state A).

The broad view of validity is needed to encompass the many factors that could affect the inferences drawn from NAEP. A broad conception of validity is also needed to deal adequately with the diversity of inferences that are based on NAEP results. Factors that might affect the validity of one inference (say, conclusions about the magnitude of cross-sectional differences among racial/ethnic groups in specific content areas) may have far less importance in evaluating the validity of another (say, conclusions about trends in achievement over time).

It is easy to lose sight of the broader validity issues when dealing with the technical details of studies designed to answer narrower questions and that sometimes require highly detailed and technical analyses. Though individually less interesting to those concerned with global conclusions, the components that are the foci of individual studies are, as was recently suggested by Crooks, Kane, & Cohen (1995), analogous to the links in a chain. A broken or weak link can seriously undermine the validity of key interpretations of the results. Hence, it is important to consider each of the components in reaching an integrated evaluation of the validity of inferences based on assessment results.

Because validity depends on the "*adequacy* and *appropriateness* of *inferences* and *actions*" based on assessment results (Messick, 1989, p. 13, emphasis in the original) it is important to be begin with a consideration of the broad purposes of NAEP and the types of inferences that it is intended to support as well as possible actions that might be based on it. Therefore, we turn to a brief discussion of the purposes of NAEP before moving on to a consideration of the findings and implications of the studies conducted by the TRP.

Purposes of NAEP

Although the first administration of NAEP took place in 1969, the idea for NAEP began six years earlier when Ralph Tyler sent a draft memorandum to then Commissioner of Education Francis Keppel in which Tyler outlined his ideas for developing dependable data that could be used to track the improvements in education (Hazlett, 1973). Shortly thereafter Commissioner Keppel approached the president of the Carnegie Corporation, John Gardner, in the fall of 1963 seeking support to develop the assessment. A small grant was awarded almost immediately to hold two conferences to consider the idea of an assessment that would provide information about the achievement of the nation's students. Ralph Tyler was asked to help plan the conferences and subsequently became the chairperson of the Exploratory Committee on the Assessment of Progress of Education that, with the assistance of its Technical Advisory Committee, chaired by John Tukey, formulated the essential characteristics of NAEP, developed its basic design features, and helped garner the necessary political and financial support during the 1964-1968 period.

Early Defining Characteristics of NAEP

Tyler's influential role in shaping NAEP, articulating its purposes, and generating support for the idea was evident even at the first conference, held in December 1963. As described by Greenbaum et al. (1977), his 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)

Although considerable elaboration of those basic ideas took place during the following five 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 remained remarkably intact as a blueprint for NAEP throughout the years that NAEP was conducted by the Education Commission of the States (1969-1982). And some of the key ideas (e.g., 1, 2 and 4) remain in effect even today. Feature 5, the assessment of age cohorts rather than grade levels, remained in force until reporting by both grade and age was started with the 1984 assessment (the first assessment with ETS as the main contractor). Although reporting by grade level has replaced age-level reporting for the main assessments, the latter is still used for the long-term trend reports (we will say more about that when we consider the implications of TRP analyses of NAEP trend reports).

Purposes and Uses of NAEP

The above early defining characteristics of NAEP deal 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 policy makers, educators, and the public with better information about student achievement. Frank Womer (1970), the first staff director of NAEP, made this clear in one of the NAEP publications, entitled "What is National Assessment." According to Womer: "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). Other than the explicit removal of authority to assess attitudes, this statement remains consistent with current intent for NAEP, which is ". . . to improve the effectiveness of our Nation's schools by making objective information about student performance in selected learning areas available to policy makers at the national, regional, State, and local levels" (P.L. 100-297, Sec. 3402).

Of course, it is possible to have agreement about the global purpose while disagreeing sharply about the nature of the information needed or the link between the information and conclusions about actions needed to improve education. Some of the long-standing issues in these regards include the level at which data will be aggregated (nation, specific subpopulations of students, state, district, school, or below) and the degree to which NAEP should serve as a unobtrusive monitor of what is, as a lever of change, or as a mechanism for uncovering causes of educational problems.

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 argued forcefully that better information was needed to make wise decisions about policies and the allocation of resources, arguing, 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. (Tyler, 1966a, 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 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 21st century), create tensions. The validity studies conducted by the TRP obviously cannot resolve such tensions. Nor can they be expected to provide answers to questions about which specific purposes should be given priority for NAEP. They can provide a starting point for evaluating the degree to which the current design provides a trustworthy basis for particular types of inferences and suggest possible improvements in NAEP if results are to be used and interpreted in ways consistent with particular purposes.

Summary of Key Themes

Beginning with the assumption that NAEP is intended to contribute to the improvement of education through the provision of "fair and accurate" information about student achievement, there remain a host of specific issues regarding the design, implementation, and uses of NAEP that will best serve this global purpose. Among the central issues are the following topics and associated questions that were addressed by the TRP:

1. Level of summarization. How can achievement in a given subject be most validly summarized? Are the accuracy and utility of the results enhanced by the use of a single global score for each subject (e.g., mathematics) or by the use of multiple scores (e.g., algebra, geometry, numbers and operations)? This issue was addressed in research conducted by the TRP on the dimensionality of NAEP mathematics assessments (Abedi, 1994; Muthén, Khoo, & Goff, 1994).

2. Motivation. Do NAEP results provide accurate information about what students know and are able to do? Or, do the results give a misleadingly low indication of student achievement because students do not put forth their best effort because they know that the results have no direct consequences for them or their schools? A series of studies was undertaken to determine the degree to which results would be affected by changes in the stakes of the assessment and/or increased student motivation on the assessment (Kiplinger & Linn, 1992; 1995/1996; O'Neil, Sugrue, Abedi, Baker, & Golan, 1992; O'Neil, Sugrue, & Baker, 1995/1996).

3. National, state, and local reporting. What is the validity of state-by-state reporting and comparisons based on NAEP results? Can state or local assessments be validly linked to NAEP results? Can NAEP results be validly linked to international assessments? The first of these questions is the central concern of the National Academy of Education Panel on the Trial State Assessment (National Academy of Education, 1992; 1993b) and was therefore not addressed by the TRP. Questions regarding linking that have implications for the use of NAEP by states and local districts, however, were investigated in one of the TRP studies (Linn & Kiplinger, 1994a, 1994b).

4. Students at risk of low achievement. How adequate is NAEP for providing information about the achievement of students who are most at risk of low achievement? Are the social context measures in NAEP adequate for this

purpose? Does NAEP provide fair and accurate measurement of achievement of identifiable groups of students who are at risk? Abedi, Lord, and Plummer (1994) addressed one aspect of the last of these questions, that is, the influence of linguistic complexity on the fairness and accuracy of assessments in mathematics for students from non-English speaking backgrounds. Berends, Koretz, and Harris (1995, forthcoming) investigated the adequacy of NAEP social context measures for characterizing students who are at risk of low achievement.

5. Student background measures. NAEP is required to "include information on special groups, including, whenever feasible, information collected, cross-tabulated, analyzed, and reported by sex, race or ethnicity and socioeconomic status." How valid are the measures of these student background characteristics, particularly the measures of socioeconomic status? Are there better measures of these characteristics that could be used? Are there other, policy-relevant, social-context measures that are not available in NAEP that account for differences in performance of the special groups identified in the NAEP legislation? These questions were investigated by Berends, Koretz, and Lewis (1994) and Berends, Koretz, and Harris (forthcoming).

6. Adequacy of long-term trends. How adequate are the long-term trend assessments for identifying changes in the relative performance, particularly for racial/ethnic population groups? Are the estimates based on the long-term trend substantially different than they would be if the trend assessment more closely mirrored the main assessment? These questions were addressed by the TRP in a study conducted by Barron and Koretz (1994, forthcoming).

7. Data quality. How adequate are the data obtained by NAEP? The TRP addressed a specific concern that inadequate time for students to complete assessments might lead to unacceptably high omit rates and large numbers of students not reaching items near the end of assessment blocks and thereby degrade the validity of the assessments (Koretz, Lewis, Skewes-Cox, & Burstein, 1992).

8. Measures of instructional experiences of students. How useful are measures of the instructional experiences obtained from teacher and student reports in accounting for differences in student achievement? Analytical aspects of this question were addressed by Muthén et al. (1995).

9. Reporting and interpreting results. What is the validity of interpretations of the NAGB achievement levels and NAEP anchor points? How accurate are the interpretations of NAEP results by policy makers and educators? Two investigations of achievement levels were undertaken by the TRP (see Burstein et al., 1995/1996; Burstein et al., 1993; Linn, Koretz, Baker, & Burstein, 1991). Those reports are not discussed here because the first report focused on initial effort with 1990 achievement levels in mathematics and those levels were subsequently reset for the 1990 assessment and reset again for the 1992 assessment and the second report focused on only one aspect of 1992 effort in mathematics. For more comprehensive evaluations of the achievement levels see the National Academy of Education (1993a), U.S. General Accounting Office (1993) and the American College Testing Program (1993) for a description of the level setting, and Lissitz and Bourque (1995) for a discussion and references to responses to the evaluations. In addition to the achievement level evaluations, however, the TRP conducted research investigating the accuracy of interpretations of NAEP reports by policy makers and educators (Hambleton & Slater, 1995) and of the anchor points and achievement levels by the media (Koretz & Deibert, 1993, 1995/1996).

Major Study Results and Conclusions

Results of the TRP studies related to each of the above issues are briefly described. These descriptions are followed by some general conclusions and recommendations for NAEP.

Level of Summarization

During the first fifteen years of NAEP, results were reported on an exercise-by-exercise basis. Summaries of average percent correct on subsets of items were also used for some purposes, but it was not until the 1984 assessment, the first conducted by ETS, that scaled scores based on an item-response theory model were introduced. The scaled scores have been the primary basis of reporting since 1984. Anchor points and achievement levels are used in an effort to give greater meaning to the results, but they are themselves tied to the NAEP scale.

In addition to the composite or overall proficiency scale scores, content-specific scores are also reported. For example, in mathematics, scores for the content areas of numbers and operations; measurement; geometry; data analysis,

statistics and probability; algebra and functions; and estimation are reported in addition to the overall mathematics scores. In most accounts, however, emphasis is given to the overall scores.

How many and which scales are long-standing questions in the assessment. The number and the nature of dimensions have potentially important implications for uses and interpretations of NAEP results. A single global index has the appeal of simplicity. Multiple dimensions may be useful, however, for tracking trends that vary by content area or by characteristics of the measures (e.g., old content vs. new, or factual knowledge vs. higher order reasoning and problem-solving skills). Although separate content scales can be defined on a variety of bases there is a substantial body of research suggesting that the separate scales are highly correlated (e.g., Carlson & Jirele, 1992; Zwick, 1987)—so highly, in the opinion of some, that the overall score is sufficient. Two studies conducted by the TRP shed new light on the question of dimensionality.

Abedi (1994) investigated the issue of dimensionality in the NAEP mathematics subscale scores in relation to students' instructional and non-instructional background variables using data from the main 1990 and 1992 assessments. Consistent with findings of others (e.g., Allen, 1990; Carlson & Jirele, 1992; and Zwick, 1987), for the total group of students there are very high correlations among the subscales. Correlations among subscale scores are lower, however, when computed for student subgroups formed on the basis of background variables.

The relevance of background variables to the question of dimensionality was made more evident in the work of Muthén et al. (1994) who used grouping variables of traditional importance to NAEP (e.g., gender, ethnicity) in multidimensional structural models. They demonstrated that this approach is more sensitive to deviations from unidimensionality than traditional factor analytic approaches that are commonly applied. They not only identified several statistically significant dimensions in addition to a general mathematics achievement factor, but found that there are differential population group differences on the secondary dimensions.

For an assessment such as NAEP where the emphasis is on population group differences—both based on background characteristics such as gender, race/ethnicity, and socioeconomic status, that are called for in the NAEP

legislation, and from one assessment to the next when looking at trends—the sensitivity of scales to group differences is of critical importance. Thus, it is important to note that even when correlations among factors are as high as they are for the NAEP content factors, differential population group differences can be identified on the various factors. Furthermore, as Muthén et al. (1994) suggest, investigations of "subgroups' differences with respect to specific factors may lead to a more 'instructionally sensitive' way to analyze achievement data" (p. 30). The latter suggestion is discussed below in connection with a subsequent study conducted for the TRP by Muthén and his colleagues concerning the analysis of measures of instructional experiences of students.

Motivation

NAEP is expected to provide information about what students know and are able to do in selected subjects. This presupposes that students participating in NAEP take the assessment seriously and try to do their best in responding to the exercises. The fundamental assumption that students put forth a reasonable effort in responding to NAEP has been called into question on several occasions, however. Shanker (1990), for example, noted that "one of the most frequently offered theories about low NAEP scores is that kids know the tests don't count." NAEP is clearly a low-stakes test for individual students. No individual student scores are reported on NAEP, and students are informed of this fact. This low-stakes character of the assessment could, as Shanker and others have suggested, result in lower performance by students than they are capable of demonstrating as the result of lack of effort to do their best. Quite simply,

> if students know that what they do on a test doesn't matter, they may decide it's not worth their while to put forth any effort. And it could be that this explains the low level of achievement we have seen on NAEP examinations. (Shanker, 1990)

The TRP conducted two major lines of research to investigate the degree to which NAEP scores may underestimate student achievement due to lack of students' motivation to put forth the effort needed to do their best. The first line of research investigated how performance on NAEP items would be affected if the items were administered to students in the context of a state assessment that had higher stakes than NAEP. The second line of research involved experimental manipulations of administration conditions designed to enhance student motivation to perform.

In the state-embedded study (Kiplinger & Linn, 1992; 1995/1996), a block of NAEP mathematics items was embedded in a state assessment that is used for state and local school accountability purposes. Although the stakes for individual students are not as high as when test results are used to make grade-to-grade promotion or graduation decisions, previous research has shown that simply reporting scores at the individual school level raises the stakes of testing for teachers and students. Hence, it was reasoned that the higher stakes associated with the state testing would lead to greater effort on the part of students when NAEP items were embedded in the state test than when they were presented in a regular NAEP administration. A small, but statistically significant, effect in the hypothesized direction was found for only one of the two subsets of items administered. Although it is possible that bigger effects might be found if rewards and sanctions for individual students were dependent on test results, the relatively small effect for only one of two subsets of items for a testing program with substantial stakes for schools suggests that NAEP results are not as depressed as the result of poor student motivation to perform as some have suggested.

The latter conclusion is made much stronger by the complementary results of the experimental studies of motivation effects (O'Neil et al., 1992, 1995/1996). Student focus groups were used to identify a range of administrative conditions that students believed would be likely to motivate greater effort on NAEP. Several of these conditions were then pilot tested, and the most promising of those conditions were then used to administer NAEP mathematics items to randomly assigned groups of students at Grades 8 and 12. The three experimental conditions used at both Grades 8 and 12 in the main study were (a) financial reward, (b) competition, and (c) personal accomplishment. At Grade 12, a fourth experimental condition, the offer of a certificate of accomplishment, was also used. Each of the experimental conditions was compared to the standard NAEP administration condition. Significant improvement in performance was found only for the financial reward condition ($1.00 for each correct item) on easy items at Grade 8. Although this indicates that NAEP may be underestimating what Grade 8 students are capable of doing given sufficient incentive to some degree, the effect was relatively small (effect size = .20) and, when coupled with the nonsignificant differences for other motivating conditions, suggests that NAEP results do not seriously understate student performance due to the low-stakes nature of the assessment.

National, State, and Local Reporting

The expansion of NAEP to allow state-by-state reporting was identified by Alexander and James (1987) as the "single most important change recommended by the Study Group" (p. 11). That change was made possible with the passing of the 1988 reauthorization of NAEP, and state-by-state comparisons have been a part of the last three assessments in two subjects (1990, Grade 8 mathematics; 1992, Grades 4 and 8 mathematics and Grade 4 reading; and 1994, Grade 4 reading). As noted above, the evaluation of the Trial State Assessment is being conducted by the National Academy of Education and is beyond the scope of this report. It is worth noting here, however, that the large number of states that have participated in each of the state-by-state assessments to date attests to the high level of interest that states have in this use of NAEP. It is also worth noting that scarce resources have limited the number of grades and subjects that could be included in the state-by-state assessment.

The interest in comparisons using NAEP other than those for the nation as a whole or for large regions of the country is not limited to states. There is an interest on the part of some local districts in obtaining district-level results. Comparisons of NAEP results to results of other countries by linking to international assessments are of interest. The latter possibility requires some type of linking procedure since the administration of NAEP in the various other countries for which comparisons are desired is not feasible. Linking is also of considerable interest to states and districts even if they are able to participate in actual NAEP assessments from time to time in particular grades and subjects because comparisons may also be desired for years in between assessments for a given grade and subject and in other subjects.

Requirements for several types of linking have been discussed by Linn (1993) and by Mislevy (1992). The most straightforward and cost-effective type of linking relies on comparisons of distributions for comparable groups of students on two assessments (e.g., NAEP and a state assessment or NAEP and an international assessment). The viability of this approach was investigated in a TRP study conducted by Linn and Kiplinger (1994a, 1994b). State tests were linked to NAEP using 1990 data in four states, and the adequacy of the linking was evaluated in terms of actual NAEP results in 1992 and results estimated from state test data using the 1990 linking function. The results indicate that such linking provides a reasonably accurate estimate for average performance in the state, but is not

sufficiently trustworthy for use in estimated performance for the highest or lowest achieving students within the state. This conclusion is supported by results reported by Erickan (1993) and in comparisons of two approaches to linking NAEP and the assessments used in the International Assessment of Educational Progress (Beaton & Gonzalez, 1993; Pashley & Phillips, 1993). Subsequent work (Bloxom, Pashley, Nicewander, & Yan, 1995; Williams, Billeaud, Davis, Thissen, & Sanford, 1995) suggests that more expensive approaches involving the administration of parts or all of both assessments to be linked may be required to achieve adequate linking.

Students at Risk of Low Achievement

From its inception, the performance of students who are at risk of low achievement has been of special interest for NAEP. Tyler talked about the need "to provide the public with dependable information to help in the understanding of educational problems and needs and to guide in efforts to develop sound public policy regarding education" (Tyler, 1966b, p. 1). He also indicated that NAEP should assess "the levels of learning of the least educated, average, and most educated groups in the society" (Greenbaum et al., 1977, p. 10). Based on their analysis of the early summaries of the original meetings regarding NAEP, Greenbaum et al. identified the first major objective of NAEP as follows: "To obtain meaningful national data on the strengths and weaknesses of American education (by locating deficiencies and inequalities in particular subject areas and particular subgroups of the population)" (p. 13). The continued legislative requirement to report by race or ethnicity and socioeconomic status may be seen as consistent with this emphasis.

Students who are at risk of low achievement were the focus of TRP research conducted by Berends et al. (1995). Since previous low achievement is the strongest predictor of poor achievement in the future, they focused their study on low-achieving students who were identified as those who score in the bottom quartile and decile of the achievement distribution. They then studied the strengths and weaknesses of NAEP for describing those at-risk students.

Berends et al. found that several measures of student, school, and community characteristics in NAEP are useful for describing at-risk students. Low-achieving groups of students were found to differ from the general population, for example, in terms of parents' education level, the average amount of

homework for a student's school, and the average years of school for parents of students who attend the student's school. Based on comparative analyses of other data sets, it was concluded that to achieve better differentiation and more complete description of the characteristics of at-risk students, NAEP would need to add or improve a number of current measures such as family income, family size, mother's age at birth of first child, previous retention in grade, student mobility between schools, and average family income of students attending a student's school. A major weakness of the NAEP for providing more complete descriptions of at-risk students is the lack adequate measures of community or neighborhood environments.

The implications of these findings depend on purposes of NAEP and the priority given to various purposes. The current measures appear adequate if the goal is to predict the probability that a student will be a low achiever. They are less adequate if an important purpose is the reporting of achievement of and the tracking of progress of the achievement of poor students. More adequate measures of family income and of school poverty would be needed for this purpose.

A second TRP study emphasizing at-risk students focused on students with language backgrounds other than English (Abedi et al., 1994). A linguistic analysis of NAEP mathematics items was conducted in an effort to identify linguistic features that might affect the performance of language minority students and students with limited English proficiency. Characteristics of the items, such as the familiarity/frequency of non-mathematics vocabulary, the voice of verb phrase, the length of nominals, and the use of conditional clauses, relative clauses, question phrases, and abstract or impersonal presentations, were identified and used to construct alternative versions of NAEP mathematics items.

The goal was to create modified versions of items with reduced linguistic complexity while maintaining the same mathematical requirements for each item. Although students from non-English speaking backgrounds showed a strong preference for the revised items over the original ones, there were no significant differences in the performance of students on the two variations of the items.

Student Background Measures

At a minimum NAEP is required to collect sufficient information about student background to meet the mandate that "the Commissioner shall— . . . include information on special groups including, whenever feasible, information

collected, cross-tabulated, analyzed, and reported by sex, race or ethnicity and socioeconomic status" (P.L. 103-382, Sec. 411 [1][C]). Such data need to be adequate to support the requirement that NAEP provide fair and accurate reporting of student achievement. Student background measures can also be useful, as was discussed above, for identifying and predicting student groups that are at risk of low achievement and possibly in providing a better understanding of the factors that have an impact on student achievement.

The availability and quality of NAEP measures of student background and social context variables was the focus of research conducted for the TRP by Berends et al. (1994; see also Berends et al., forthcoming). The adequacy of measures was evaluated by comparing the relationships of NAEP social context and background variables to student achievement with those found in other national data bases that have richer data on social context, including, for example, data collected from parent interviews rather than only through the type of student, teacher, and principal reports that NAEP relies upon.

The research found that NAEP is missing a number of social context measures (e.g., number of siblings, parental occupation) that have been found in other national studies to account for part of the observed performance differences between racial/ethnic groups. The reliance on student self-reports and other unreliable data sources to get information about family income and average income for a school also undermines the ability of NAEP to portray population group differences that remain after taking student and school poverty into account. Consequently, "NAEP usually overestimates the achievement differences between students who come from different population groups but similar social contexts. However, at the secondary school level, these overestimates reflect primarily the absence of important measures rather than reliance on student self-reports; for several reason, predictions based on parent and student reports are similar" (Berends et al., 1994, pp. ii-iii).

How important the limitations of NAEP measures of student background and social context are depends on the purposes and priorities of NAEP. At Grades 8 and 12, due to the particular variables at issue, the reliance on student reports does not seriously erode the prediction of population group differences in achievement. Because there is less consistency between parent and student reports of family characteristics at the fourth grade than at the two higher grade levels, reliance on student self-reports is a more serious limitation at Grade 4 than

at Grades 8 or 12. For other purposes (e.g., understanding how economic circumstances of students attending a school influence achievement), the limitations of the NAEP social context measures are more serious.

Adequacy of Long-Term Trends

As its name suggests, the tracking of progress in educational achievement is fundamental to the concept of NAEP. The basic function of monitoring progress was reaffirmed in the 1994 reauthorization: ". . . the Commissioner shall—. . . report achievement data on a basis that ensures valid and reliable trend reporting" (P.L. 103-382, Sec. 411 [1][B]). How adequately is NAEP meeting the fundamental purpose of valid and reliable reporting of trends in student achievement? This question was addressed in research conducted by Barron and Koretz (1994; forthcoming).

In considering NAEP trend reports, a distinction must be made between the main and trend assessments. In recent years the function of monitoring long-term trends has been separated from the main assessment, which is used for cross-sectional reports of current status and for reporting short-term trends. This split was the result of major changes in the content frameworks and associated assessments in recent years that brought into question the comparability of the current assessments with those that had been used in earlier years. The finding in the previously discussed reading anomaly that even seemingly minor changes in assessment procedures could result in relatively large artifactual shifts in scores on the assessment further reinforced the split. Indeed, a major conclusion that was reached in the study of the reading anomaly was that "when measuring change, do not change the measure" (Beaton, 1990, p. 165). As a consequence the long-term and main NAEP assessments are administered to separate samples of students. This allows the main assessment to introduce exercises developed to reflect new content frameworks and innovative approaches to measurement while using assessment booklets and administration procedures that are identical to ones used in the past in the long-term assessments.

The focus of the Barron and Koretz study was on the long-term trend assessment, particularly the accuracy of trend estimates for racial/ethnic population groups. They found that, because of the smaller sample size used in the long-term trend assessment in comparison to the main assessment and the lack of oversampling for minority populations, the resulting estimates for African

American and Hispanic students have large standard errors. Because of the large standard errors, the tracking of long-term trends in the achievement of racial/ethnic minority populations lacks the degree of precision required to detect potentially important changes in the performance of the these population groups.

Barron and Koretz noted that there are substantial differences between the main and long-term trend assessments. Differences between the two assessments in terms of changes in the proportion of items in various content categories, in item format, and in the cognitive processes that the items are intended to measure may produce disparate results for the two assessments.

There is, as others have noted (e.g., Beaton & Zwick, 1990; Zwick, 1992), a tension between the need to maintain continuity with the past so that change can be measured and the desire to introduce innovations and improvements in the assessment to make it consistent with the best current thinking and to make it more forward-looking. The NAEP approach of having two assessments serving these competing needs is sensible and so is the plan that would keep old, long-term trends in place until a new trend could be firmly established and tied, if feasible, to the old one. Presumably the current, main-assessment procedures would become the long-term trend in future years when the then current main assessment introduced future content frameworks and approaches to assessment (possibly including, for example, computer-administered problems and simulations). As noted by Barron and Koretz, however, "a new trend assessment will not solve several of the fundamental problems brought up in this study. For example, reliable estimates of trends for minorities will require a substantial change in sampling, one which might require reallocating resources from the main to the trend assessment" (forthcoming, p. 30).

There is a need for more extended discussion and reconsideration of the approach being used to measure long-term trends. Two issues that need attention in this regard are (a) the level of precision that should be sought in the trend results for racial/ethnic population groups, and (b) whether there are alternatives to the strict adherence to the use of identical measurement procedures in measuring long-term trends. With regard to the first of these issues the key question is the size of the standard error of minority group means that is acceptable. The related question is then one of identifying an efficient and cost-effective design that will achieve the needed level of precision.

The second issue is more complicated. It is not only the case that the definition of core subjects may change so much over the course of a couple of decades that a measure that was consistent with the definition at one point in time might be of much less interest at another. Recent experience suggests that such changes may be occurring on a much more rapid cycle. Certainly, there are minor changes in emphasis that are desired from one assessment to the next, and changes are also desired in the format of the assessment tasks. Also, as suggested by Goldstein (1983) and Zwick (1992), it cannot be assumed that keeping the items the same necessarily means that the measure is unchanged. The same items may measure different things as the result of changes in context and instruction. Thus, rather than relying only on a procedure that holds everything a constant as possible, an alternative approach to maintaining comparability over time may be the inclusion in the overall assessment design of "multiple means of checking that the scale has been preserved" (Barron & Koretz, forthcoming, p. 31).

Data Quality

The validity of any assessment of student achievement depends on the quality of the raw data provided by students. As noted above, inferences about student proficiency based on assessment results depend on the assumption that students put forth a reasonable effort on the assessment, which led to the investigations of the effects of student motivation on NAEP scores. Valid inferences about what students know and can do also depend on basic assumptions that the administration conditions allow students an adequate opportunity to respond. If students are not provided a reasonable amount of time to respond to NAEP items, for example, then low scores could give a misleading indication of what students are capable of doing when given adequate time to respond.

In the jargon of psychometrics a distinction is commonly made between speed tests and power tests. The number right on a pure speed test would simply equal the number of items responded to within the time limit. That is, given enough time it is assumed that all respondents could answer any of the items correctly on a pure speed test. The score on a pure power test, on the other hand, would not increase if the testing time was increased. In practice, almost all tests and assessments involve some combination of power and speed, but the intent is

usually to minimize the influence of one of these factors, most commonly speed, so that conclusions can be drawn about the other factor (power, i.e., what students know and can do given a reasonable opportunity and adequate time). Clearly, the type of conclusions that NAEP is intended to support require an assessment that is not overly influenced by speed.

Two types of nonresponse, omitted and not-reached items, are distinguished on NAEP. An item is considered omitted if a student fails to respond to the item but does respond to a subsequent item within the timed block of items. When a nonresponse occurs toward the end of a block and none of the subsequent items in that block are responded to, the item is classified as not-reached. Not-reached item rates have traditionally been used as one of the indicators of the speededness of a test.

Nonresponse rates on items on the 1986 mathematics assessment were high enough to raise serious concerns that the results on that assessment might have been unduly influenced by speed. In the 1986 mathematics assessment, 23% (104 of 446) of the items had not-reached rates of .20 or higher. The not-reached rates were so high that the response rate criterion for including an item in the mathematics scale had to be relaxed so that items with not-reached rates as high as .45 were included in the scaling. Even with this lenient criterion 15% (79 of 446) of the items were excluded from the NAEP mathematics scale.

Concern about the apparent speededness of the 1986 mathematics assessment led to the conduct of a TRP study of omitted and not-reached items on the 1990 assessment (Koretz et al., 1992). Although the study of nonresponse was stimulated by the high nonresponse rates in the earlier assessment, it was also deemed to be important because of changes in the format of NAEP, particularly the increase in the proportion of constructed-response items. Nonresponse rates tend to be higher on constructed-response items than on multiple-choice items. Moreover, the concern was expressed that disparate rates of nonresponse on constructed-response items across groups of students could lead to erroneous inferences about group differences in proficiency.

Koretz et al. (1992) found that the overall not-reached rates were substantially less in the 1990 mathematics assessment than they were in the 1986 assessment. At Grades 4 and 8, omit rates were modest. Differences in nonresponse for White and minority students could be only partially explained,

however, by differences in mathematics proficiency. At Grade 12, high omit rates were more common, particularly on some of the constructed-response items. The higher omit rates on those items for African American and Hispanic than for White Grade 12 students also raises concerns, particularly in light of the increasing reliance on constructed-response items by NAEP. Because of these concerns and the differences in nonresponse rates that have been noted from one assessment to another, Koretz et al. recommended that "focused monitoring and reporting of non-response patterns" (p. 22) become a routine part of NAEP technical analyses and reports.

Measures of Instructional Experiences of Students

Although NAEP is not designed to provide a basis for making causal inferences about the effects of different educational policies on student performance, NAEP is frequently used to compare the achievement of students who have different instructional experiences. Student achievement is separately reported, for example, for students attending public and private schools, and for students enrolled in academic, vocational, or general high school programs. Two nagging questions in such comparisons concern (a) the degree to which these comparisons reflect demographic differences rather than differences in instructional experiences and (b) whether NAEP is sufficiently sensitive to detect specific instructional influences in addition to global differences in overall proficiency.

A study conducted for the TRP by Muthén et al. (1995) investigated analytical aspects of both of these questions. Muthén and his colleagues began by identifying factors in the teacher questionnaire responses regarding instruction that could be used to describe differences in the instructional experiences of students participating in the Grade 8 mathematics assessment. They identified four factors: (a) an "NCTM factor," that is, schools where teachers reported that emphasis was given to aspects of mathematics stressed by the National Council of Teachers of Mathematics (1989) such as communicating mathematical ideas, appreciation of mathematics, reasoning/analysis; (b) a remedial and typical mathematics instruction factor where emphasis is on learning mathematical facts and concepts, skills and procedures, and numbers and operations; (c) an enriched classroom factor where the emphasis was given to geometry, and to

data, statistics, and probability, while little attention was given to measurement; and (d) an algebra factor (i.e., eighth-grade classes emphasizing algebra).

Using the multidimensional structural modeling procedures discussed earlier in connection to investigations of the dimensionality of NAEP mathematics assessments by Muthén and colleagues (1994), Muthén et al. (1995) investigated effects of class type on performance not only on the general mathematics achievement factor but on specific factors with student background variables and achievement on the general factor held constant. At Grade 8 they found that student-based algebra class type and teacher-reported NCTM factor had effects on algebra-specific factor. At Grade 12 they found strong effects of studying geometry-trigonometry on geometry-specific performance and of studying algebra-calculus (but also strong effect for studying trigonometry) on algebra-specific performance after controlling for the general proficiency factor and background variables. Differential effects not explained by the general factor were also found for a problem-solving factor.

As noted by Muthén et al. (1995) "the analyses point to new possibilities in terms of choice of scoring and reporting of achievement components. [Specifically], the fact that different effects are found for the content-specific factors and for the problem-solving factor than for the general factor motivates further investigations of such achievement components" (p. 46). It should be noted, however, that better and more detailed measures of the instructional experiences that students have would need to be added to NAEP in order for the analytical approach illustrated by Muthén and his colleagues to be of greater utility in the analysis and interpretation of NAEP data.

Reporting and Interpreting Results

The scaled scores used in NAEP reporting since ETS took over the operation of the assessment in 1984 have numerous important advantages, particularly for reporting trends, but those come at a substantial price: Scaled scores in themselves are not meaningful to most audiences. Is a decline of 3 points trivial or important? Lay readers have no way to know.

NAEP has used two methods for imparting intuitive meaning to scaled scores. The first entailed setting "anchor points" at arbitrarily chosen points on the distribution of scores. Statistical criteria were then used to select items that concretely illustrate performance at each of them, and these items in turn

provided the basis for verbal descriptions of each of the anchor points. The second method, started after the 1988 reauthorization of NAEP, sets three performance standards for each grade, called "achievement levels," reflecting judgments about the levels of proficiency students should show. These levels were presented with descriptions of varying detail and exemplar items in order to give them meaning to lay audiences.

Although the anchor points and achievement levels differ in important respects—most importantly, in that the anchor points reflect the ad hoc distribution of performance, while achievement levels represent judgments about what students should be able to do—both represent efforts to give meaning to scaled scores by describing and illustrating a few specific points on the scale. Little information was available, however, about the success of these efforts in communicating NAEP results accurately to lay audiences. Accordingly, Koretz and Deibert (1993) systematically reviewed the presentation of the results of the 1990 mathematics assessment in the lay print media during a seven-month period in which NCES and NAGB released publications that used both metrics.

The anchor points and achievement levels struck a responsive chord—nearly all of the articles reviewed used one or the other of these metrics in reporting primary (national- and state-level) results. Scaled scores were used much less frequently and usually in conjunction with anchor points or achievement levels. But both anchor points and achievement levels were used less frequently by the press to report secondary findings, such as sex differences and differences between population groups.

Some of the effects of the two metrics on reporting were undesirable, however. Writers usually used only the simplest of the descriptions of anchor points and achievement levels provided in NCES and NAGB publications, and they sometimes simplified these descriptions further. Although both the NCES and NAGB reports provided more substantial descriptions of the knowledge and skills of students at the points or levels, relatively few writers made use of them. The use of anchor points and achievement levels also seems to have encouraged writers to misrepresent student achievement as discontinuous—students either can or cannot do what is in the descriptions of the levels. Both of these tendencies are illustrated, for example, by a statement that students at Level 200 "know how to add."

Another misunderstanding arose because the percentage of students reaching an anchor point or achievement level is often quite different than the percentage correctly answering items used to illustrate it. This difference, which can be large, stems in part from the nature of the statistical screens used to select exemplars. To lessen the likelihood of confusing these two percentages, both the NAGB and NCES reports provided actual *p*-values (for all students and for students at the anchor points or achievement levels) for illustrative test items. The provision of *p*-values, however, had relatively little effect on the press reports and did not prevent the confusion of *p*-values with the percentage of students reaching the levels. Relatively few articles presented any illustrative items, and in the articles following the release of the NAGB and Goals Panel reports, most of those that did present percentages clearly misconstrued the percentage of students reaching the achievement levels as being the *p*-values for illustrative items.

The achievement levels (unlike the anchor points) reflect judgments about how students should perform, and different panels of judges (or different methods for setting the levels) would likely have produced different standards. Only a small minority of the articles that discussed achievement levels made any mention of the judgmental nature of the levels, and most of those did so only briefly. The implications for the robustness of the levels was not made apparent.

This study illustrates that the effectiveness of presentations of NAEP results cannot be taken for granted. Rather, ongoing research is needed to establish empirically which methods of presentation work best for specific purposes and which have unacceptably serious unintended effects.

Additional evidence regarding the need for empirical evidence regarding effective methods of presentation that are understandable to intended audiences while minimizing misinterpretations is provided by Hambleton and Slater (1995) in their interview study of 59 policy makers and educators. Hambleton and Slater found that these users had considerable difficulty with the presentation of results in the NAEP executive summary report. They documented that the important audiences that were the target of their study had a considerable amount of misunderstanding of results. The policy makers and educators were confused, for example, by: (a) average proficiency scores, (b) standard errors, (c) the use of > and < symbols to denote significant differences (increases/decreases), and (d) the use of "at or above" in describing the percentage of students who score at a given

achievement level (e.g., Basic) or higher (e.g., anywhere in the Basic, Proficient, or Advanced categories). With regard to the latter point, the respondents misinterpreted the percent "at or above" as the percent in each proficiency category and then became confused when percentages did not add to 100.

Hambleton and Slater suggest that there is a need for more user-friendly reports with considerably simplified figures and tables. Consistent with the Koretz and Deibert (1993) media study, Hambleton and Slater also make a strong case for field testing graphs, figures, and tables with target audiences.

Discussion and Conclusions

As was noted in the introduction, validity is a multifaceted concept that depends on the particular uses and interpretations of assessment results as well as on the instruments and administration conditions. NAEP is expected to serve a wide variety of purposes, and the results it produces are interpreted in manifold ways by and for a diverse array of audiences. Thus, it is meaningless to ask a single global question: "Is NAEP valid?" and expect a simple yes-or-no answer. The response must clearly be that NAEP has a high degree of validity for some particular uses and interpretations, but little validity for others.

The TRP studies described above contribute to the overall understanding of the validity of specific interpretations and help identify interpretations and uses that are suspect or areas where the assessment needs to be strengthened. For example, the studies of the effects of motivation on performance on NAEP suggest that a major potential threat to the validity of NAEP as a means of describing what students know and are able to do is far less serious than some critics have suggested. While those studies do not answer all the questions about the degree to which the lack of student effort tends to deflate scores, they do indicate that the performance would not be likely to increase noticeably as the result of modest increases in stakes or by the offering of reasonable incentives to students for better performance.

Other studies point to ways in which NAEP could be strengthened and/or raise questions about priorities and tradeoffs between serving different purposes. The analyses of the long-term trend data, for example, show that the current design needs strengthening if it is to provide a reasonably sensitive indicator of long-term changes in the relative performance of population groups identified in

the authorizing legislation and of long-standing concern for NAEP (i.e., racial/ethnic minorities, and economically disadvantaged students). While the changes in design that would be needed to strengthen the long-term trend for these purposes are relatively straightforward and easily articulated, the decision to make those changes is, of course, much more complicated and requires a consideration of the trade-offs in the use of scarce resources and policy priorities between long- and short-term trend and between long-term trend and other NAEP priorities (e.g., state-by-state reporting, number of subjects areas, inclusion of students with limited English proficiency, students in private schools, and students with handicapping conditions that require assessment adaptations or accommodations).

In a similar vein, study results provide a clear indication that better measures of some student background characteristics (e.g., family income) and of specific instructional experiences of students are needed for some potential uses of NAEP data. The results do not answer the policy questions regarding the importance of getting solid information about family income or the degree of poverty of students enrolled in a school. The legislation does call for reporting results by "socioeconomic status" where "feasible," but that does not answer the question of priority or specificity of measures that are needed. Nor does it address the question of how to weigh the desire for better measures of socioeconomic status against the cost, the potential burden of collection of family income information through procedures such as household surveys of parents, and inevitable trade-offs with other desired activities.

The instructional experiences suggestion raises more fundamental policy questions about the degree to which NAEP should provide the basis for analyses of instructional factors that are related to and possibly explain student proficiency. As the brief recounting of some of the early history in the introduction indicates, the degree to which NAEP should serve this type of purpose has been an issue since the early planning days of NAEP. Frank Womer's (1970) early cautions about the limits of NAEP staked out a position that would clearly put concerns for better measures of instructional experiences beyond the scope of NAEP. But many interpretations of NAEP invite implicit causal explanations involving instructional experiences (e.g., instructional emphasis placed on "literature-based reading," on "phonics," or on "whole language" [Mullis, Campbell, & Farstrup, 1993, p. 132], emphasis on "knowing science facts and terminology"

or on "developing skills in laboratory techniques" [Jones, Mullis, Raizen, Weiss, & Weston, 1992, p. 93], or for that matter, enrollment in a vocational education program).

Other TRP study results involve less sweeping policy questions. The recommendation for making analyses of omit and not-reached rates a part of the routine data analyses involves relatively minor and relatively inexpensive additions to the current technical work of the NAEP contractor yet would provide an important addition to the set of procedures used to monitor and maintain data quality.

Although at one level the results of the studies of dimensionality primarily involve technical issues, the findings have more sweeping implications for reporting and interpreting NAEP results. They suggest alternate ways of reporting data that might enable tracking changes in student performance in terms of specific segments of content domains that go undetected in global composite scores. Such a change could have important implications for the overall utility of NAEP.

Earlier we referred to the analogy that Crooks et al. (1995) drew between aspects of validity and the links of a chain. The utility of a chain depends on the strength of the links. In an analogous fashion the validity of conclusions about what students know and are able to do depends on administration conditions that lead students to put forth a reasonable effort and that allow them sufficient time to attempt the tasks that are presented. Thus, analyses of seemingly mundane characteristics such as omit and not-reached rates can have critical implications for the overall dependability of the chain that supports important inferences.

The TRP studies of reporting point to an often ignored aspect of validity: the adequacy of different approaches for presenting assessment results. Since validity is a characteristic of inferences based on scores, the effectiveness with which assessment reports encourage supportable inferences and discourage incorrect inferences is an essential link in the validity chain. Even presentations that seem clear to the writer may inadvertently mislead readers, in particular the lay readers who comprise some of NAEP's most important audiences. The TRP studies underscore the need for continuing empirical evaluation of the adequacy of reporting methods.

Although the focus of this paper has been narrowly limited to the studies conducted by the TRP, the studies obviously do not stand alone. There is a large and growing body of research conducted by the main NAEP contractor, through grants and contracts awarded by NCES, and by independent secondary analysts using the NAEP data files that contribute to the overall understanding of the validity of NAEP for its many uses and interpretations. Such research, including but not limited to that by the TRP, helps provide NAEP with a firm foundation.

There is now a substantial body of research bearing on the validity of NAEP for various purposes, but the need for further studies continues. Any number of important questions remain unaddressed. For example, there has been increasing pressure to use NAEP to provide information pertaining at the level of schools rather than the level of individual students. The current sampling system is poorly designed for that purpose; to change it to provide better school information would be reasonably straightforward but might exact a substantial price in terms of greater margins of error for other important statistics. How should this conflict be resolved? Moreover, the use of NAEP is continuously evolving, and the demands and expectations that confront it are expanding. This will create new questions of validity, new trade-offs, and the need for further research.

References

Abedi, J. (1994). *Achievement dimensionality, Section A* (Tech. Rep., Draft, May). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., & Plummer, J. R. (1994). *Language background as a variable in NAEP mathematics performance* (Tech. Rep., Draft, November). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Alexander, L., & James, H. T. (Eds.). (1987). *Improving the assessment of student achievement: The nation's report card*. Washington, DC: National Academy of Education.

Allen, N. L. (1992). Data analysis for the science assessment. In E. G. Johnson & N. L. Allen (Eds.), *The NAEP 1990 Technical Report* (Tech. Rep. No. 20-TR-20, pp. 275-302). Princeton, NJ: Educational Testing Service.

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.

Barron, S. I., & Koretz, D. M. (1994). *An evaluation of the robustness of the NAEP trend lines for racial/ethnic subgroups* (Tech. Rep., December). Santa Monica, CA: RAND.

Barron, S. I., & Koretz, D. (forthcoming). NAEP trend lines for racial/ethnic subgroups. *Educational Assessment*.

Beaton, A. E. (1990). Epilogue. In A. E. Beaton & R. Zwick (Eds.), *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly* (Tech. Rep. No. 17-TR-21, pp. 165-168). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Beaton, A. E., & Gonzalez, E. J. (1993). Comparing the NAEP Trial State Assessment with the IAEP international results. In *The National Academy of Education. Setting performance standards for student achievement: Background studies*. Stanford, CA: The National Academy of Education.

Beaton, A, E., & Zwick, R. (Eds.). (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly* (Tech. Rep. No. 17-TR-21). Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Beaton, A. E., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics*, *17*, 95-109.

Berends, M., Koretz, D., & Harris, E. (1995). *Identifying students at risk of low achievement in NAEP and NELS* (Draft, June). Santa Monica, CA: RAND.

Berends, M., Koretz, D., & Harris, E. (forthcoming). Minority test scores and social context. *Educational Assessment*.

Berends, M., Koretz, D., & Lewis, E. (1994). *Measuring racial and ethnic test score differences: Can the NAEP account for dissimilarities in social context?* (Draft, March). Santa Monica, CA: RAND.

Bloxom, B., Pashley, P. J., Nicewander, W. A., & Yan, D. (1995). Linking to a large-scale assessment: An empirical evaluation. *Journal of Educational and Behavioral Statistics*, *20*, 1-26.

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

Carlson, J., & Jirele, T. (1992). *Dimensionality of 1990 NAEP mathematics data*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147-171). Urbana: University of Illinois Press.

Crooks, T., Kane, M., & Cohen, A. (1995). *Threats to the valid use of assessments*. Unpublished manuscript, University of Otago, Dunedin, New Zealand.

Erickan, K. (1993). *Predicting NAEP*. Unpublished manuscript, CTB Macmillan/ McGraw-Hill, Monterey, CA.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, *20*, 269-277.

Greenbaum, W., Garet, M. S., & Solomon, E. (1977). *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 the state-*

level NAEP comparisons (NCES Tech. Rep. CS 89-499). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Hambleton, R. K., & Slater, S. C. (1995). *Are NAEP executive summary reports understandable to policy makers and educators?* (Tech. Rep., Draft, May). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Hazlett, J. A. (1973). *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 of Kansas.

Hearings before the Select Committee on Equal Educational Opportunity of the U.S. Senate. (1971). *Ninety-Second Congress, First Session on Equal Educational Opportunity Part 22—Education Information*. Washington, DC. December 1, 2, 3.

Jones, L. R., Mullis, I. V. S., Raizen, S. A., Weiss, I. R., & Weston, E. A. (1992). *The 1990 science report card: NAEP's assessment of fourth, eighth, and twelfth graders*. Washington, DC: National Center for Education Statistics.

Kiplinger, V. L., & Linn, R. L. (1992). *Raising the stakes of test administration: The impact on student performance on NAEP* (CSE Tech. Rep. No. 360). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Kiplinger, V. L., & Linn, R. L. (1995/1996). Raising the stakes of test administration: The impact on student performance on NAEP. *Educational Assessment, 3,* 111-333.

Koretz, D. M., & Deibert, E. (1993). *Interpretations of National Assessment of Educational Progress (NAEP) anchor points and achievement levels by the print media*. Santa Monica, CA: RAND.

Koretz, D. M., & Deibert, E. (1995/1996). Setting standards and interpreting achievement: A cautionary tale from the National Assessment of Educational Progress. *Educational Assessment, 3,* 53-81.

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1992). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress* (Tech. Rep.). Santa Monica, CA: RAND.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*, 83-102.

Linn, R. L., & Kiplinger, V. L. (1994a). *Linking statewide tests to the National Assessment of Educational Progress: Stability of results* (CSE Tech. Rep. No.

375). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. L., & Kiplinger, V. L. (1994b). Linking statewide tests to the National Assessment of Educational Progress: Stability of results. *Applied Measurement in Education*, *8*, 135-155.

Linn, R. L., Koretz, D. M., Baker, E. L., & Burstein, L. (1991). *The validity and credibility of the achievement levels for the 1990 National Assessment of Educational Progress in mathematics* (CSE Tech. Rep. No. 330). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Lissitz, R. W., & Bourque, M. L. (1995). Reporting NAEP results using standards. *Educational Measurement: Issues and Practice*, *14*(2), 14-23, 31.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Mislevy, R. J. (1992). *Linking educational assessments: Conceptual issues, methods, and prospects*. Princeton, NJ: Educational Testing Service, Policy Information Center.

Mullis, I. V. S., Campbell, J. R., & Farstrup, A. E. (1993). *NAEP 1992 reading report card for the nation and the states: Data from the national and trial state assessments*. Washington, DC: National Center for Education Statistics.

Muthén, B. O., Huang, L., Jo, B., Khoo, S., Goff, G. N., Novak, J., & Shih, J. (1995). *Opportunity-to-learn effects on achievement: Analytical aspects* (CSE Tech. Rep. No. 407). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Muthén, B. O., Khoo, S., & Goff, G. N. (1994). *Multidimensional description of subgroup differences in mathematics achievement data from the 1992 National Assessment of Educational Progress* (Tech. Rep., January). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

National Academy of Education. (1992). *Assessing student achievement in the states. The first report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment: 1990 trial state assessment*. Stanford, CA: National Academy of Education, Stanford University.

National Academy of Education. (1993a). *Setting performance standards for student achievement. A report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels*. Stanford, CA: National Academy of Education, Stanford University.

National Academy of Education. (1993b). *The trial state assessment: Prospects and realities. The third report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment: 1992 trial state assessment*. Stanford, CA: National Academy of Education, Stanford University.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1992). *Final report of experimental studies on motivation and NAEP test performance* (CSE Tech. Rep.). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

O'Neil, H. F., Jr., Sugrue, B., & Baker, E. L. (1995/1996). Effects of motivational interventions on NAEP mathematics performance. *Educational Assessment, 3,* 135-157.

Pashley, P. J., & Phillips, G. W. (1993). *Toward world-class standards: A research study linking international and national assessments*. Princeton, NJ: Educational Testing Service.

Shanker, A. (1990, July 29). How much do our kids really know? Raising the stakes on NAEP. *The New York Times*.

Staff of the National Assessment of Educational Progress. (1977). Response of the National Assessment of Educational Progress. In W. Greenbaum, M. S. Garet, & E. R. Solomon, *Measuring educational progress* (pp. 193-229). New York: McGraw-Hill.

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

Tyler, R. W. (1966b). The objectives and plans for a National Assessment of Educational Progress. *Journal of Educational Measurement*, *3*, 1-4.

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

Williams, V. S. L., Billeaud, K., Davis, L. A., Thissen, D., & Sanford, E. (1995). *Projecting to the NAEP scale: Results from the North Carolina end of grade testing program* (Tech. Rep. No. 34.). Research Triangle Park, NC: National Institute of Statistical Sciences.

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*, no. 2, 118-120.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement*, *24*, 293-308.

Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Statistics*, *17*, 205-218.