Reporting the Results of the National Assessment of Educational Progress

Richard M. Jaeger
Center for Educational Research and Evaluation
University of North Carolina at Greensboro[1]

Commissioned by the NAEP Validity Studies (NVS) Panel
September 1998

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

---

1 This paper was prepared while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences at Stanford University. Partial support of the Spencer Foundation under Grant Number 199400132 is gratefully acknowledged.

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

Panel Members:

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Patricia Dabbs
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
1791 Arastradero Road
PO Box 1113
Palo Alto, CA 94302
Phone: 650/ 493-3550
Fax: 650/ 858-0958

Contents

List of Tables

Introduction

Since its first administration in 1969, those responsible for the National Assessment of Educational Progress (NAEP) have grappled with the problem of reporting and disseminating its results in forms that reach intended audiences, are understood by potential users, and promote valid interpretation. In its earliest years, the agency that operated NAEP employed a professional journalist with responsibility for fashioning reports of NAEP results that were, at once, interesting and understandable to the public and technically accurate. The task was daunting and little success was claimed. As has been noted elsewhere (Jaeger 1996), reporting to the public on a project with the scope and complexity of NAEP is extremely difficult, both in the selection of an appropriate reporting vehicle and in the choice of form and format for reported information:

> Carefully crafted technical reports, no matter how accurate and appropriately guarded in conveying fine nuances of conclusion and interpretation, will rarely see the light of day beyond the offices of measurement specialists and a small cadre of assessment policymakers. The press craves provocative information and simplification, while those who create assessment reports strive for cautious communication and interpretive accuracy. The objectives and needs of these groups appear to be fundamentally inconsistent (1).

By law and in fact, NAEP serves a variety of audiences, each with differing needs for its information, differing interests in its findings, and differing sophistication in interpreting its results. Among these audiences are elected officials and civil servants at federal and state levels (members of Congress, the President and his staff, the Secretary of Education, other members of the Cabinet, professionals in the U.S. Department of Education and in other federal agencies, governors, state legislators, and professionals in state education agencies), education policymakers and executives in local education agencies (school board members, school superintendents and their executive staff members), educators in schools (principals and teachers), educational researchers, members of the general public (parents with children in school, taxpayers, members of public advocacy groups), and members of the press (broadly defined to include newspapers, television reporters, and radio reporters).

Reporting vehicles and reports best suited for some of these audiences will not likely be best for others. For example, technical detail on sampling of students, analysis of data, and precision of findings, which will be demanded by educational researchers and technically sophisticated measurement personnel in state departments of education, are not likely to be of interest to policymakers and executives at federal, state, or local levels. It is clear that no single report on NAEP results will meet the needs of its entire constituency. This has been recognized by the National Center for Education Statistics (NCES), and has resulted in the publication of a variety of reports on NAEP and its outcomes, which for the 1998 assessment will include *NAEP Report Cards*, containing

the results of a single assessment and intended for policymakers; *Update Reports*, focusing on a single issue and intended for parents and other members of the public; *Instructional Reports*, containing assessment materials and intended for educators; *State Reports*, containing the results of a NAEP state assessment and intended for state education executives; *Cross-State Data Compendia*, intended for state education executives and educational researchers; *Trend Reports*, documenting long-term trends in students' achievement and intended for educational researchers and policy analysts; *Focused Reports*, addressing important policy issues and intended for educational policy analysts and researchers; *Almanacs*, containing NAEP data for secondary research; and *Technical Reports*, documenting the procedures used in conducting the assessment and intended for educational researchers and psychometricians (NCES 1997). NCES clearly has endeavored to provide its audiences with a wide variety of sources on the National Assessment. Little is known, however, about the effectiveness of these various reports in providing NAEP's constituencies with needed information in forms that are understandable and useful.

Communicating the results of a major assessment program such as NAEP presents distinct challenges. The breadth of the audiences to be reached, their differing interests, their differing access to various dissemination vehicles, and their vastly differing technical backgrounds makes effective communication especially difficult. Furthermore, the challenge of effective communication is multifaceted. It is not enough to know how various NAEP audiences might be reached. It is also essential to understand the kinds of information they desire, the forms of information that might be useful to them, and the formats in which information might be understandable and applicable.

This paper explores the ways the National Assessment results might be communicated to its varied constituencies. It contains three main sections. The first section begins by exploring the forms in which NAEP's fundamental findings on student achievement have been conveyed, and concludes with some proposals that have been advanced for alternate reporting models. The second section summarizes some additional considerations with implications for any new research agenda on NAEP reporting. The third section builds on the information provided to that point in order to suggest a detailed program of research on how best to report and portray NAEP's findings.

History of NAEP Reporting

Characterization of NAEP Achievement Results—How Have NAEP Achievement Results Been Summarized?

Student achievement results from NAEP have been summarized in a variety of ways. These are well described in a report titled "Interpreting NAEP Scales" (Phillips et al. 1993). In the original conception of NAEP, results were reported in terms of students' collective performances, exercise-by-exercise. The proportion of tested students who answered an exercise correctly (called a p-value) was reported for each NAEP exercise, overall, and for major subgroups, including those classified by region, gender, size and type of community, education level of the students' parents, and race/ethnicity. This approach to reporting is consistent with the vision held for NAEP by its principal architect (Tyler 1966) and is illustrated in the first NAEP report on science achievement (Education Commission of the States 1970).

Although reporting results by item embodies an appealing simplicity and clarity, the sheer volume of reported statistics made it difficult for users to integrate and understand students' achievement in a comprehensive way. According to Phillips et al. (1993):

> ...the early mode of reporting many items together with their p-values highlighted a problem that persists today—how to communicate a comprehensive view of NAEP findings in a brief and accurate manner. When reporting the first wave of assessments across curriculum areas, it became clear that for the most part, educators, policymakers, and the public did not have the time to study and assimilate the voluminous item-by-item results. The problem for NAEP audiences trying to understand the results became particularly acute when considering findings across a variety of subject areas (10).

In 1977, Mullis, Oldefendt, and Phillips attempted to address the issue of excessive detail by reporting the characteristics of NAEP exercises in a given subject matter field that had p-values within prescribed ranges. An example of this strategy is given in Phillips et al. (1993) for grade 4 students in the 1992 NAEP Mathematics Assessment:

Many fourth-graders (more than two-thirds) can:

- Add and subtract two- and three-digit whole numbers when regrouping is required;
- Recognize numbers when they are written out;
- Identify instruments and units for measuring length and weight; and
- Recognize simple shapes and patterns.

Some fourth-graders (approximately 33 percent to 67 percent) can:

- Solve one-step word problems, including some division problems with remainders;
- Work with information in simple graphs, tables, and pictographs;
- Round numbers and recognize common fractions; and
- Substitute a number for "□" in a simple number sentence.

Few fourth-graders (less than one-third) can:

- Solve multistep word problems, even those requiring only addition and subtraction;
- Perform computations with fractions;
- Solve simple problems related to area, perimeter, or angles; and
- Explain their reasoning through writing, giving examples, or drawing diagrams.

Although the efficacy of this mode of summarization and reporting does not appear to have been examined empirically, one can readily posit several shortcomings. First, there is no assurance that the skill characteristics reported within any range of p-values is representative of the tested skills associated with that range. Second, the volatility of item p-values as a function of minor changes in item format and content is well known. Hence, the generalizability of the statements across sets of items that fall within a description such as "Perform computations with fractions" is suspect. Third, the p-value ranges for which skills have been summarized are quite broad. In particular, a range that varies from one-third of students to two-thirds of students includes what some would regard as reasonable success and what others would regard as abject failure. Finally, the skills reported within a given p-value range are quite diverse, and do not obviously lend themselves to ready conceptual summarization in terms of a curriculum framework.

A third approach to characterizing NAEP achievement results, used from the time the need to report achievement trends first arose and, for special assessments, into the late-1980s, involved reporting the average p-values associated with sets of items in a portion of the NAEP content domain for students at various age or grade levels. A recent example of this type of reporting can be found in Martinez and Mead (1988), a report on the first National Assessment of students' computer competence. That assessment provided achievement results for students in grades 3, 7, and 11.

In addition to reporting percent correct values by item, average percent correct scores and associated standard errors were reported for items in such categories as "knowledge of

computer technology," "understanding of computer applications," "knowledge of computer programming," and "overall computer competence." Six sets of percentages were reported for items in each of these categories: items that were used exclusively in grade 3, items that were used exclusively in grade 7, items that were used exclusively in grade 11, items that were used in grades 3 and 7, items that were used in grades 7 and 11, and items that were used in all three grades.

It is difficult to make comparisons across grade levels using the average p-value metric because curriculum typically differs materially in different grades, even within the kinds of topical categories used here for reporting. It is not surprising, therefore, that substantial differences between average p-values were found for students at a single grade level across items in any of the categories. Unfortunately, this result casts doubt on the generalizability of the overall findings. For example, in the "Knowledge of computer programming" category, students in grade 11 had an average percent correct of 27.2 on items used exclusively in that grade, but an average percent correct of 38.8 on items used with students at all three grade levels.

As Phillips et al. (1993) pointed out, the average p-value metric also was problematic when trends were reported across assessments. Average p-values are trustworthy only for subsets of items that are common to successive assessments, but dealing with this problem by restricting trend comparisons only to the common items gave rise to other problems of representativeness. In addition, when trends were reported across three assessments (e.g., NAEP 1978), separate analyses were needed for items that were common to the first two assessments, for items that were common to the second and third assessments, and for items that were common to all three assessments. Finally, average p-values only provide information on the central tendency of students' collective achievements and offer no information on other distributional features. The result is a somewhat confusing and less-than-complete portrayal of students' performances.

Since 1984, NAEP has used item response theory (IRT) and Bayesian statistical methods to produce scaled National Assessment scores for populations and subpopulations of students. The current NAEP design and analysis procedures were introduced by Messick, Beaton, and Lord (1983) and have been documented in detail by Beaton and Johnson (1992), Mislevy, Johnson, and Muraki (1992), and Mislevy et al. (1992). The use of scaled scores for reporting students' NAEP performances affords a number of important advantages compared to reliance on p-values for individual items or exercises, but imposes some difficulties as well. Perhaps the most important advantage is the ability to compare the performances of subpopulations of students within the same assessment or the performances of populations and subpopulations across assessments, even though the groups of students compared did not all complete the same set of test items. IRT provides comparable estimates of students' abilities and comparable estimates of the difficulties of items even when subgroups of students responded to different test items.

A second advantage that scaling affords is estimates of distributions of student achievement for various populations and subpopulations, including estimates of percentile ranks. When scaling procedures are used, it is no longer necessary to restrict descriptions of groups' performances to reports on their mean performances.

A third advantage, since the NAEP performances of populations and subpopulations are described in terms of a continuous variable, is the ability to relate group scores on a performance scale to a set of continuous and discrete background variables. It is now possible, for example, to estimate the correlation between NAEP scale scores and indicators of students' socio-economic status.

One disadvantage of portraying NAEP results in terms of scaled scores is that the scaling metric is arbitrary. Until very recently, all of the NAEP scales used values that ranged from a minimum score of zero to a maximum score of 500 across the three grade levels tested (grades 4, 8, and 12 in the most recent assessments). In the base year, each scale had a mean of 250.5 and a standard deviation of 50 points (Linn and Dunbar 1992). Although the use of an arbitrary scale metric is common in educational testing (for example, the Scholastic Assessment Tests and the Graduate Record Examinations make use of subtest scales that, when introduced, had a mean of 500 and a standard deviation of 100), users have difficulty determining the importance of given scale score differences until such scales have been used for a number of years. For example, is a five-point difference in mean scale score from one NAEP assessment to the next an important difference or a trivial difference in terms of educational importance? Not surprisingly, this question still has not been answered satisfactorily for NAEP. As a result, NAEP reports focus on the statistical significance of the differences between mean NAEP scores for various subpopulations or on the statistical significance of changes in mean NAEP scores, across assessments, for a given population or subpopulation.

Since statistical significance is sample-size dependent and the samples used to estimate mean NAEP scaled scores are typically large, differences that are substantively unimportant often will be identified as statistically significant. In contrast, despite the disadvantages of reporting NAEP results in terms of the percentage of examinees who answered a given item correctly, it is not difficult to decide whether a 5 percent gain in the percentage of students who answered a given item correctly is or is not important. Therefore, the meaning of the reporting metric is clear from the outset, and there is no generalization beyond the item for which the p-values are reported.

Another difficulty with the zero-to-500 cross-grade scale is that those who interpret NAEP results in the popular press often do not realize that the scale is defined across all three NAEP grade levels. It is not unusual to see results for students at grade 4 or grade 8 interpreted as though 500 is the expected maximum score for students in that grade.

Furthermore, as noted by Linn and Dunbar (1992, 186), the zero-to-500 scale for NAEP results does not facilitate ready interpretation of intermediate scores, such as 200 or 300. In response to this problem, Phillips et al. (1993) identified two approaches that have been used at different times in the history of the National Assessment. The first, termed "item mapping," identifies for a large number of NAEP items the scale score at

which 80 percent of students answered the item correctly. That is, an item was placed on the NAEP scale at the point where the conditional probability of answering the item correctly was 0.80.

This approach was first used with the reports of the 1985 NAEP Literacy Assessment of Young Adults (Kirsch and Jungeblut 1986). Items were described in a shorthand that conveyed their central features. In illustrating this form of reporting for the 1992 NAEP Mathematics Assessment, Philips et al. (1993) showed that fourth-grade students scoring at a scale value of 156 had an 80 percent chance of correctly answering an item that required them to multiply 3×405 using a calculator, students scoring at a scale value of 178 had an 80 percent chance of correctly answering an item that required them to add two 3-digit numbers, and students scoring at a scale value of 301 had an 80 percent chance of answering correctly a word problem that involved multiplication of 3 by 11/3. A total of 30 fourth-grade items were thus located on the NAEP scale within the scale interval 150 to 301. Item locations were shown graphically, with the NAEP scale displayed as a vertical bar with scale values of 150, 200, 250, and 300 prominently identified. The graph also contained the information that the average mathematics scaled score of fourth-graders was 218 with an associated standard error of 0.7, that 98 percent of fourth-graders scored at or above 150, that 72 percent scored at or above 200, and that 17 percent scored at or above 250. Similar graphs were produced for eighth-graders and for twelfth-graders.

A second approach to characterizing NAEP results on the zero-to-500 scale involves what has been termed "scale anchoring." Introduced in the report on the 1983–84 Reading Assessment and used as late as the 1992 Mathematics Assessment (Mullis et al. 1993), scale anchoring is a strategy for describing the meaning of students' knowledge and skills at designated positions on the NAEP scale; in this case, the scale values 200, 250, 300 and 350. The process began by identifying items that were likely to be answered correctly by students who scored within a 25-point band immediately surrounding a selected scale value, and that were less likely to be answered correctly by students at the next lower scale value. For example, to meet the requirement of an anchor item at level 250, an item had to:

1. Be answered correctly by at least 65 percent of students with scale scores immediately surrounding 250;

2. Be answered correctly by at least 30 percent fewer students with scale scores immediately surrounding 200;

3. Be answered incorrectly by at least 50 percent of students with scale scores immediately surrounding 200; and

4. Have been attempted by at least 100 students with scale scores immediately surrounding each of 200 and 250.

The second and third criteria were not used for anchor items at the lowest anchor level, Level 200. Once items that satisfied the statistical criteria had been identified, groups of mathematicians and mathematics educators were assembled to succinctly describe the knowledge and skill demands of items that anchored at a given NAEP scale value. The resulting descriptions are illustrated by the Level 200 description prepared for the 1990 and 1992 Mathematics Assessments:

Level 200	Addition and Subtraction, and Simple Problem Solving with Whole Numbers
	Students at this level can identify solutions to one-step word problems involving addition or subtraction. They can add and subtract whole numbers in most situations, and when a calculator is available, they can multiply and divide. They are able to select the largest whole number from a set of numbers in the thousands, and can match the verbal and symbolic names for numbers. Students demonstrated familiarity with length and weight, by selecting appropriate instruments and units to measure these attributes. They are able to recognize some basic properties of two-dimensional geometric figures as well as the names of standard examples of these figures. They can extend simple patterns.

Once the descriptors for Levels 200, 250, 300, and 350 had been created, NAEP results were reported in terms of the percent of students at each of grades 4, 8, and 12 whose scaled scores equaled or exceeded these levels. Results for the 1990 and 1992 Mathematics Assessments were compared by first computing the percentage of students, by grade, whose scaled scores equaled or exceeded each of the four anchor levels and then determining the statistical significance of differences between corresponding percentages for 1990 and 1992. Of course, average scaled score values also were reported for each grade and year, as were the results of tests of the statistical significance of differences between corresponding averages. Beaton and Allen (1992) provide a full description of the scale anchoring method.

Unfortunately, as documented by Linn and Dunbar (1992), prominent education reporters misinterpreted the meaning of NAEP anchor items that were used as exemplars of students' knowledge and abilities at the various scale values. One statistic reported for each exemplar item was the percentage of students with scores near a given scale value (e.g., Level 250) who could answer the item correctly. However, this conditional percentage was incorrectly interpreted as the unconditional percentage of students who

could answer the item correctly. Forsyth (1991) gave an example of this error in an article by Adler (1990) that appeared in *Newsweek*. In that article, Adler stated for a reported anchor item at Level 300, "8th-grade problem: 15.9 percent of 13 year-olds and 51.1 percent of 17 year-olds answered correctly questions like this (18)." As noted by Linn and Dunbar (1992, 186), the item had not been administered to 13 year-olds and was answered correctly by 74.1 percent of 17 year-olds. Another similar error was made by Shanker (1990). He confused the percentage of students above a given anchor level with the percentage of students who answered given anchor items correctly. Rothman (1991) also made a similar error when reporting NAEP results in *Education Week*. His conclusions suggested that the percentage of students who could answer an exemplar item for a given anchor level was equal to the percentage of students who earned scale scores at or above that level. Linn and Dunbar (1992) concluded from errors such as these that:

> The correct interpretation of anchor items may be too complicated for their intended purpose of providing greater meaning for the scale… Despite the desirability of integrating educational research with public policy, some separation of the two seems necessary in the context of public reports of NAEP results. The confusion that has surrounded the interpretation of anchor points represents a major threat to the validity of NAEP. How to display the data from NAEP to the public in a way that clearly differentiates the data themselves from the public policy statements that are based on the data remains an issue of critical importance to the overall validity of NAEP (190–191).

This uncharacteristically pessimistic conclusion by Linn and Dunbar highlights the difficulty that has attended reporting of NAEP results since its inception. There continues to be an underlying tension between the quest for accuracy on the part of NCES personnel who have the responsibility for upholding the standards of a major federal statistical agency, the seemingly inexhaustible demand of the popular press for quick, simple, and problematic news about the state of public education in the United States (Jaeger 1992), and the public's need for straightforward reports that impose minimal demand for interpretation of statistical findings.

In 1988, when the Congress reauthorized NAEP, it created an independent policymaking group called the National Assessment Governing Board (NAGB). As part of the authorizing legislation for NAGB, the Board was charged with developing "appropriate achievement goals for each…grade in each subject area to be tested under the National Assessment" (Public Law 100–297). NAGB acted on this charge by establishing what it terms "achievement levels" for NAEP. These levels, designated *Basic*, *Proficient*, and *Advanced* for each tested grade, depart from the NAEP tradition of characterizing students' actual NAEP performances. Rather, they were designed to indicate the levels of performance students *should* exhibit.

Beginning with the 1990 assessment in mathematics, NAEP results have been reported in terms of the percentages of students within a grade, nationwide and by

subpopulation, whose scores result in their classification as below *Basic*, *Basic* or above, *Proficient* or above, or *Advanced*. Although scaled score means have been reported as well, greater emphasis has been placed on the achievement-levels results, and the press has focused almost exclusively on the achievement-level results in their recent coverage of NAEP (Jaeger 1996, April).

The NAEP achievement levels differ from their predecessor anchor levels in several important ways. First, the achievement levels define judgmental specifications of *desired* student performance rather than empirically-derived positions on a scale of students' actual NAEP performances. Second, achievement levels are developed separately for each grade level in which NAEP is administered, while anchor levels are defined on a scale that spans all three grades. Third, the anchor levels define points on the NAEP scale, and the achievement levels define intervals of NAEP performance. Although the NAEP achievement levels have been embraced enthusiastically by the press and by state testing directors (DeVito 1997; Hawkins 1995), preliminary evidence suggests that they do not consistently foster valid interpretation (Hambleton and Slater 1995).

As noted by Kane (1994), any judgmental standard-setting process involves two steps. The first step results in the definition of a performance standard—a verbal specification of what examinees should know and be able to do if they are to be classified as, for example, *Proficient*. The second step results in a cut score—a score on the scale of the test or assessment that purports to identify examinees who just barely satisfy the performance standard. For NAEP, the first step begins with a set of brief statements adopted by NAGB as a matter of policy, as definitions of achievement that is *Basic*, *Proficient*, or *Advanced*. As an example, the core NAGB definition of *Proficient* is: "This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter." [2] Historically, NAGB's definitions of achievement levels included predictive components; for example, the definition of *Proficient* used with the 1992 assessments was "This central level represents solid academic performance for each grade tested—4, 8, and 12. It reflects a consensus that students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling" (Phillips et al. 1993, 38). These kinds of predictive definitions invite counter-examples, and ready questioning of their validity. For this reason, the current definitions which exclude predictive statements, although subject to broad interpretation, are likely to be more defensible.

The first step continues with the amplification of NAGB's brief descriptors through reference to the NAEP content frame for the relevant assessment and grade level. For many NAEP assessments, at least the first draft of these amplified descriptors was prepared by the curriculum specialists who developed the content frame, but had no further responsibilities in defining the cut scores. For more recent assessments, the

2 From the NAGB web site <http://amcom.aspensys.com/nagb/abtnagb.html≠levels> 11/19/97).

content-specific performance standards were further modified by a panel of persons with responsibility for completing the second step.

An important component of the NAEP performance standards is a set of exemplar exercises associated with each achievement level for each grade. These exercises are selected by a judgment panel on the basis of their perceived quality, some norm-referenced criteria (at least half the examinees at a given level must answer the exercise correctly if the exercise is to be used as an exemplar of that level; also, increasing percentages of examinees at higher levels must answer the exercise correctly), consistency with the NAEP content frame, and appropriateness for the grade level for which it was considered. In recent reports on the results of NAEP assessments, three exemplar exercises are shown for each achievement level within each grade (c.f., Reese et al. 1997).

The second step—defining cut scores on the NAEP scale corresponding to *Basic*, *Proficient*, and *Advanced* achievement for students in each tested grade—is completed by a panel composed in large part of teachers and other educators, but including a substantial percentage of non-educators. The procedure used to elicit the judgments of panel members is an iterative variant of the Angoff method (Angoff 1971; Jaeger 1989). The validity of the cut scores that have resulted from this method has been a topic of ongoing debate that is, as of this writing, unresolved (Cizek 1993; Jones 1997; Kane 1993; Linn 1996; Shepard et al. 1993; Stufflebeam, Jaeger, and Scriven 1991; U.S. General Accounting Office 1993). At issue, among other questions, is whether the method imposes a judgment task that is virtually impossible to address; i.e., predicting the difficulties of test items for a subset of hypothetical examinees with abilities that are consistent with a conceived performance standard (Chang 1996; DeMauro 1995; Impara and Plake 1996; Quereshi and Fisher 1977; Taube and Newman 1996; Thorndike 1980; Wheeler 1991). The larger question, which has yet to be extensively explored, is the validity of the interpretations by users of NAEP results when results are reported in this metric. However, preliminary evidence on this issue is not encouraging. Bourque and Garrison (1991) found that the NAEP achievement levels often were interpreted in the press as statements of what students can do, rather than what they should do. Jaeger (1996) found in press reports based on the NAEP achievement levels, a number of outlandish inferences that went far beyond their defensible meaning.

Thus the methods used to summarize students' collective NAEP achievement have varied throughout the decades, and none has met with unbridled acclaim. For the most part, criticism has been speculative and research on the effectiveness of any of the methods has been sparse and inconclusive.

Alternative Representations—Recent Proposals for NAEP Reporting

Over a year ago, NAGB began a dialogue on how NAEP could be redesigned so as to better achieve its goals within its financial constraints. This resulted in a redesign report issued by NAGB (1996) and the commissioning of an analysis of the technical feasibility of the proposed redesign (Forsyth et al. 1996). In their design feasibility report, Forsyth et al. discussed an interesting alternative for structuring NAEP and for

reporting its results, called the "market-basket" approach (for further elaboration, see Mislevy 1996). With market basket reporting, NAEP performance reports would include the release of a representative sample of the items and exercises used in an assessment, together with their scoring rubrics. In one of three alternative definitions of a market basket discussed by Mislevy (1996), the released items would constitute one of a number of psychometrically parallel collections actually administered to individual students. The advantage of this definition of a market basket is the possibility of reporting students' performances in an observed-score metric that would avoid the need for the kinds of sophisticated statistical manipulations of data required by the current IRT scaling. For example, it would be possible to simply report the percentage of possible score points earned by the average student in a population or subpopulation. Seeing the collection of items referenced by the percentage, it is hoped that NAEP users could more readily understand the meaning and the limitations of NAEP scores. When discussing the shortcomings of the current IRT-based reporting, Mislevy insightfully noted:

> When the tasks fade into the background as discussion increasingly revolves around the more abstract term 'mathematics proficiency,' it is too easy to forget about the aspects of 'mathematics proficiency' that cannot be captured under standard NAEP data-collection activities (18–19).

As Mislevy indicated, the market-basket approach to reporting would make clear precisely the kinds of skills NAEP assessed as well as the skills it did not assess under the label mathematics proficiency or achievement. Whether NAEP users would be sufficiently adept to discern NAEP's strengths and limitations by reviewing a sample of its items and exercises, and whether they would be sufficiently interested in doing so, is an empirical question that certainly would have to be investigated, should the market-basket approach be adopted.

A related proposal for NAEP reporting was put forth by Bock (1997) and amplified in Bock, Thissen, and Zimoski (1997). It is similar to Mislevy's proposal in that the principal reporting metric would be the percentage of a domain of items and exercises that a group of examinees, on average, could answer correctly. Bock posits the superiority of percent-correct reporting for promoting public understanding:

> Reporting to the media and the general public in terms of scale scores…presents problems in giving concrete meaning to the results and making clear what the units of the scale represent…A further difficulty with scale scores is that they are metric quantities rather than counts. Quantities, with their particular units, are problematic in reports intended to inform the public debate and policy, where results typically have to be expressed in unitless percentages to be understood (84).

Bock's proposal differs from Mislevy's in its methodology and in the mechanism that would be used to inform users about the domain. Bock suggested that the NAEP domain in a subject area be defined operationally by a very large set of items and exercises that

would be systematically adopted or constructed to represent every facet of a NAEP content framework. The set of items would then be partitioned into two randomly-equivalent halves and one half would be released to the public prior to the administration of the relevant NAEP assessment. The other half would be partitioned into many parts to compose test forms of convenient length. Prior to release of the half-domain to the public, the entire domain would be pilot tested using a matrix sampling design similar to that currently used in NAEP assessment, for the purpose of estimating IRT item parameters. Then when the assessment was administered operationally, some form of IRT estimation (Bock illustrated the superior precision of a Bayesian IRT estimation procedure) would be used to estimate each examinee's domain score in the percent-correct metric and, in turn, to estimate the average domain score of subpopulations of interest. Because IRT estimation would be used, it would not be necessary to administer representative or randomly parallel test forms to individual examinees.

Both the Mislevy (1996) and the Bock (1997) proposals are variants of an idea first proposed by Ebel (1962) in an article titled "Content Standard Test Scores." Like Mislevy and Bock, Ebel felt that test scores would gain meaning for technically unsophisticated users if the scores referenced a well-defined body of content that was made manifest for test users.

In a memorandum to NAGB and its staff concerning the market-basket proposal for NAEP, Haertel (1997, August) endorsed a display of results put forth by Ebel (his Table 2) as a particularly useful way of linking performance on a disclosed market-basket of test items to performance on the NAEP content domain and scale. Ebel suggested a disclosed test composed of ten items that would be linked to the score scale of an underlying domain-referenced assessment by showing, for each of a number of "true" scale scores, the portion of 100 examinees who would earn each possible observed score on the disclosed test. In Ebel's example, there was a monotonic relationship between the true scale score on the entire domain and the central tendency of the distribution of observed scores on the disclosed test. The same would be true for NAEP. As Haertel noted, this kind of display would have the advantage of indicating the uncertainty associated with any inference that linked an examinee's observed score on the disclosed market-basket and his or her true score on the underlying NAEP assessment. Although appealing to measurement specialists as much for the modesty of its claims as for its communication value, it is doubtful that NAEP results reported in terms of Ebel's Table 2 would be satisfying to policymakers and educators. First, they probably would want a linkage that went in the other direction: "Given a score on the disclosed market-basket, what score could I expect on the NAEP scale?" Second, they probably would want to eliminate the uncertainty inherent in the relationship. They would likely want to know the NAEP score they could expect on the basis of a market-basket score rather than a distribution of possible NAEP scores. Here again, the tension between the desire for technical soundness on the part of measurement specialists and the desire for simplicity on the part of policymakers and educators come into conflict.

In October 1997, NAGB released a report and an eight-page companion document intended for non-technical audiences (such as parents and citizens) on the results of the 1996 National Assessment in Science. These reports differed in format and content from earlier NAEP reports issued by NCES in several respects. First, they focused on the percentages of students nationwide and in various subpopulations, including states and jurisdictions that participated in the NAEP state assessments, whose performance resulted in their classification in the four achievement-level categories—below *Basic*, *Basic*, *Proficient*, or *Advanced*. Second, NAEP results were shown in formats that might make them more readily understood by lay readers. For example, standard errors of estimates were provided in separate tables rather than being shown beside corresponding estimates in a single table.[3] The eight-page companion document emphasized achievement-level definitions and exemplar exercises from NAEP, and contained very few data tables. The tables that were included provided percentages of students, by grade level and race/ethnicity or by grade level and gender, whose performances placed them in each NAEP achievement-level category. No standard errors were given for these estimates. Similar data for states and jurisdictions that participated in the 1996 NAEP state assessment in science were shown graphically.

These achievement-level reports might be prototypic of future reports issued by NAGB. Their comprehensibility and utility for various purposes and audiences certainly should be investigated.

Other Considerations with Implications for Research on NAEP Reporting

Prior Research on NAEP Reporting

As Hambleton and Slater (1995) indicated, far greater attention has been paid to the design, development, scoring, and scaling of NAEP assessments and to their core psychometric properties than to the reporting of NAEP's findings to its various audiences. Some attention has been given to the preferences of some NAEP users for several forms of NAEP results and to the information desires of some NAEP users, but little is known about the most effective ways to report NAEP's findings on student achievement in the nation and the states or to disseminate those findings. Ward (1980) surveyed potential NAEP users at federal, state, and local levels to secure their judgments on the kinds of information they needed from NAEP. She found that many groups suggested more extensive analysis and interpretation of NAEP results. Few of Ward's findings were heartening. Most of the teachers surveyed were not familiar with NAEP and saw little usefulness in its findings. Some respondents suggested that NAEP results be disseminated

3 The convention of displaying standard errors in separate tables rather than the main data table also has been adopted by NCES for NAEP reports beginning with the 1996 assessment.

through teachers' journals so as to increase the likelihood that teachers would become aware of its existence. Although greater familiarity with NAEP was evident among respondents at federal and state levels, many complained that NAEP results were of little utility for a variety of reasons. Notably, a number of users called for the establishment of performance standards that could be used to gauge the quality of student achievement.

Koretz and Deibert (1993) analyzed the accuracy and reasonableness of statements in the print media concerning students' performances on the 1990 National Assessment in Mathematics. In that assessment, results were reported in terms of anchor levels, with their accompanying exemplar exercises, and in terms of achievement levels, then newly introduced by NAGB. The authors found serious errors in the interpretation of results presented in both metrics. Writers used highly simplified versions of the definitions of anchor and achievement levels presented in NAEP reports. They then erroneously represented students' abilities in terms of dichotomies—as though students either could do or could not do what the anchor levels or achievement levels described. As noted earlier in a study by Linn and Dunbar (1992), Koretz and Deibert found that education writers confused the percentage of students who had reached an anchor point with the percentage of students who could correctly answer items used to illustrate the anchor point. In addition, achievement levels were falsely represented as definitions of students' current capability rather than judgmentally-based descriptions of desired performance. In sum, this study indicated that NAEP results frequently were misinterpreted in the popular press.

Hawkins (1995) studied, among other issues, the perceived utility of the reporting formats used with the 1992 NAEP Trial State Assessment (TSA) in Reading (grade 4) and Mathematics (grades 4 and 8) and preferences of state education personnel for several types of NAEP reporting formats. She conducted telephone interview surveys of state assessment directors and curriculum specialists, and achieved very high rates of participation from those whose states participated in the TSA.

Both assessment directors and curriculum specialists indicated that they preferred having NAEP results reported in terms of achievement levels rather than anchor levels. They found achievement levels easier to understand, more relevant for communicating the state of education, and more likely to impact education policy. Respondents endorsed the use of NAEP reporting formats that were "easy to interpret, user friendly, had good layouts, and were good for use with general audiences." They also endorsed greater use of graphs and color in NAEP reports.

State assessment directors strongly endorsed reporting of NAEP results in terms of achievement levels and noted that their own state assessments were moving to adopt similar reporting metrics. Many commented that NAEP reports were released too late and were overly long and complex. They recommended the production of short reports that are "user friendly" and designed for particular audiences, such as reading curriculum coordinators or teachers. They also suggested a program of research through which various NAEP audiences could recommend the content and format of reports that were tailored to their needs.

Hambleton and Slater (1995) conducted an interview survey of 59 state-level educators and policymakers to determine the degree to which they could understand and correctly interpret the results presented in the *Executive Summary of the NAEP 1992 Mathematics Report Card for the Nation and the States*. These researchers conducted face-to-face interviews with their research subjects, during which the subjects read and responded to questions about brief sections of the report. The authors' findings were disturbing. They concluded:

> Despite the fact that the interviewees tried hard to understand the report, we found that many of them made fundamental mistakes. Nearly all were generally able to understand the text in the report, though many would have liked to see more descriptive information (e.g., definitions of measurement and statistical jargon and concrete examples). The problems in understanding the text involved the use of statistical jargon…The tables were more problematic than the text for most of the interviewees. Although most were able to get a general feeling of what the data in the tables meant, many mistakes were made when we asked the interviewees specific questions (14).

Hambleton and Slater go on to point out numerous misinterpretations and overinterpretations of statistics in tables and graphs made by their interviewees. Their findings are particularly disheartening in light of their report that almost half their interviewees had completed more than one statistics course and only 27 percent had no formal education in statistics. In addition, almost two-thirds of their interviewees had read NAEP reports prior to the research study and nine-tenths had previous knowledge of NAEP. These results clearly indicate the need for additional research on NAEP reporting that includes comparative analyses of the efficacy and interpretability of alternative reporting forms and formats.

Jaeger (1996) reported the results of a content analysis of newspaper articles on the 1990 NAEP TSA in Grade 8 Mathematics and the initial release of the *First Look* report for the 1994 NAEP TSA in Grade 4 Reading. He found that reports of both assessments were often interpreted erroneously. In particular, differences between students' mean scores in various states were interpreted as real even though they were not statistically reliable, achievement differences among states were haphazardly interpreted as indicators of the comparative quality of their respective schools, and a variety of governmental officials at federal and state levels offered causal explanations of students' achievement that were of dubious validity. Not unexpectedly, governmental officials often cited a host of school factors to explain poor results, while education officials cited societal factors. Perhaps a classic example of making much of nothing was the reaction of North Carolina's then Superintendent of Public Instruction when he learned that the average score of his state's fourth-graders had increased by two scale-score points on the NAEP TSA in Reading from 1992 to 1994: "I was just so happy to have good news." He then attributed the gain to the state's revamped curriculum and testing system, and "five years

of improvements in the Scholastic Assessment Test." How improvement in the assessment scores of a self-selected sample of high school students caused better reading scores for the state's fourth-graders was never explained.

Jaeger (1996) suggested several modifications to NAEP reporting that might reduce the frequency with which results were misinterpreted, including more effective disaggregation of results, reporting in ways that emphasized the uncertainty surrounding summary statistics, more frequent use of simple graphical displays, and providing explicit examples of erroneous interpretations and overgeneralizations of results. Whether these strategies would achieve desired goals is currently unknown, and that is why much of the research proposed in the following section of this paper is sorely needed.

DeVito (1997) collected judgments from state testing directors on ways that the NAEP TSA could be improved. Twenty-eight assessment directors responded to a questionnaire he circulated, and commented on reporting issues among many others. DeVito also conducted a focus group at a major professional meeting to secure additional judgment data. Among other findings, DeVito learned that state assessment directors made little use of findings from national NAEP, but did use results for their own state. They compared their state's results to those of other states, used item formats to suggest changes in their own statewide assessments, and interpreted their state's results in light of state-adopted curricula.

Some Literature on Reporting of Test Results

The measurement literature contains several interesting papers on how test results should be organized and reported. Aschbacher and Herman (1991) draw heavily on related psychological literature and research in business and marketing in suggesting ways to format and organize tables and graphs so as to enhance comprehension. Generalization of these findings to achievement test reports is somewhat an act of faith, but the recommendations they make are certainly sensible. The suggestions made by Hambleton and Slater (1995) for simplification of graphs and tables are plausible, and provide a good basis for structuring alternative data presentation formats in future research studies, even though they haven't been validated with real consumers of test reports. They call for narrative explanation combined with graphical display of results.

In a lead article in the spring 1997 issue of the *Journal of Educational and Behavioral Statistics*, Howard Wainer illustrates a number of clever ways in which tabular and graphical data displays can be formatted so as to emphasize important results and eliminate the unimportant. Although Wainer's suggestions have not been validated, it seems plausible that they will result in improved communication. His suggestions could be used to great advantage in the design of studies on the interpretability of alternative NAEP reporting formats.

NCES as a Federal Statistical Agency: Implications for NAEP Reporting and Dissemination

Any research on the reporting and dissemination of NAEP results must be conducted with due recognition of the role of NCES (the federal agency responsible for operation of the National Assessment program and reporting its results) as a federal statistical agency. The principal role of NCES is to report to the Congress on the status of American education. In fulfilling its role, NCES must uphold strict standards of data quality, must provide a permanent archive of information and data on the status of American education, and must refrain from interpretations of the results of any of its myriad surveys, including NAEP, that are not strictly warranted by the data at hand.

A principal responsibility of any contractor that supports NCES in its operation and reporting of results of the National Assessment is the preparation of archival reports that fully and accurately document the methods used to collect data, the results that were found, and the precision of those results. NCES conducts a strict review and adjudication in which all reports prepared by its contractors, including reports on the National Assessment, are evaluated against criteria that reflect the agency's role as a federal statistical agency. This review procedure and the NCES criteria, may well limit the flexibility of contractors in drawing broad interpretations of findings, and perhaps, in formatting and displaying results in ways that would be most accessible to lay audiences.

The program of research suggested below assumes that reporting and dissemination of NAEP results must extend beyond the principal, archival responsibility of NCES if the full potential of NAEP to better inform the public, educators, and policymakers at all levels of government about the achievement of the nation's students is to be realized.

A Program of Research on Reporting and Dissemination of NAEP Findings

A program of research on reporting and disseminating NAEP results might be characterized in terms of three intersecting dimensions:

- The research questions to be asked;
- The audiences to whom the questions should be addressed; and
- The strategies through which the questions should be pursued.

These dimensions will be examined in turn, and will be followed by a suggested integration that yields a series of proposals for research studies.

The Research Questions

Three fundamental research questions might undergird a program of inquiry on reporting and disseminating NAEP results. First, in what form should NAEP results be reported? Second, how should NAEP results be displayed? Third, how should NAEP results be disseminated? Each of these questions could be interpreted in a multiplicity of ways.

In what form should NAEP results be reported? This question does not refer principally to choices concerning the disaggregation of NAEP results across subpopulations or to choices of correlates of students' achievement that should be reported. It is understood that NAEP results will consist of data on the collective assessment performances of students in various populations and subpopulations and of contrasts among those performances within an assessment and across assessments. Here, however, the question refers to the form in which students' collective performances on the NAEP assessment are summarized. Examples of alternatives have been reviewed earlier in the section documenting the history of NAEP reporting, and include such choices as percent-correct statistics for individual exercises, average percent correct for exercises of a particular kind, average scale scores, percentages of students at or above achievement levels, etc.

Five subquestions warrant investigation here:

1. What do various NAEP audiences find to be of interest?
2. What do various NAEP audiences find to be useful?
3. What do various NAEP audiences understand?
4. What can various NAEP audiences validly interpret?
5. Among alternatives, what do various NAEP audiences prefer?

As an instance of subquestion 3, "What do various NAEP audiences understand?", there are answers to such questions as: "How will members of a particular audience comprehend the meaning of the results reported in table 2.2 of the *NAEP 1996 Mathematics Report Card for the Nation and the States*, which indicates a 4-point change in average score for Minnesota from 1992 to 1996, with the footnote 'Indicates that the change since 1992 in average scale score is significant at a 5 percent level of significance using a multiple comparison procedure based on 39 jurisdictions (excluding the nation).'?" For subquestion 4, "What can various audiences validly interpret?", an example would be: "Will members of a particular audience refrain from making the inference that the statistically significant gain in average grade 4 mathematics score for Minnesota proves that the quality of Minnesota's mathematics instruction for fourth-graders is better in 1996 than it was in 1992?". The distinction between

"understanding" and "valid interpretation" is here intended to be one of comprehension versus drawing defensible inferences on the basis of comprehended results.

As noted in the brief review presented earlier, some of these questions have been examined for some audiences, but no comprehensive study of NAEP reporting has explored all of them. It must also be recognized that these five subquestions might produce conflicting, rather than mutually reinforcing, answers. There might well be important differences between the information and reporting formats that users prefer and those that result in valid interpretations of NAEP findings. In particular, as James Chromy (the designer of the original NAEP sampling plan) has observed, policymakers and lay readers eschew uncertainty, but uncertainty is a fundamental component of the valid interpretation of results of any sample survey, including NAEP.

How should NAEP results be displayed? This question refers to the format used to display NAEP results rather than to the content of NAEP assessment reports. That is, interest is in choices among various forms of tabular summary, various forms of graphical summary, various forms of narrative summary, and combinations of these forms of portrayal. Narrative alternatives include verbal restatement of results shown in tables or graphs, policy-grounded interpretations of findings, and cautions on incorrect or inappropriate interpretations. The same five subquestions listed earlier also warrant investigation here:

1. What do various NAEP audiences find to be of interest?
2. What do various NAEP audiences find to be useful?
3. What do various NAEP audiences understand?
4. What can various NAEP audiences validly interpret?
5. Among alternatives, what do various NAEP audiences prefer?

How should NAEP results be disseminated? At present, the principal modes of dissemination of NAEP results are through a variety of print reports, through a listing of those reports on the Internet, and through downloadable images of selected reports from a web site maintained by NCES. A more complete set of alternatives, portrayed in the following list, would include current forms of dissemination as well as additional modes:

- **Print Reports**
  Full NAEP reports on a single assessment
  NAEP summary reports on a single assessment
  NAEP briefings on specific issues (e.g., the socio-economic correlates of NAEP performance)
  Reports on trends across assessments

- **World Wide Web**
  Full NAEP reports on a single assessment
  NAEP summary reports on a single assessment
  NAEP briefings on specific issues (e.g., the socio-economic correlates of NAEP performance; regional differences in NAEP performance)
  Reports on trends across assessments

- **Data Archives for Secondary Analysis and Interpretation**
  Raw data tapes
  Web-based data archives
  Web-based summary tables

- **Public Print Media**
  Press releases
  Magazine articles—general circulation
  Magazine articles—specialized circulation (e.g., *School Administrator*)

- **Television**
  Press releases for television
  Requested interviews on news/discussion programs
  Videotapes for professional use

- **Radio**
  Press releases for radio
  Requested interviews on news/discussion programs
  Audio tapes for professional use

A number of researchable questions are associated with the best choices among dissemination vehicles for NAEP results:

1. What vehicles are accessible to various NAEP audiences?

2. What vehicles are regularly used by various NAEP audiences?

3. What vehicles are preferred by various NAEP audiences?

4. What types of information can feasibly be disseminated through various vehicles?

Audiences for NAEP Results

NAEP has the potential for serving a wide variety of audiences with varied interests in and needs for information on student achievement. Among these audiences are the following:

- **Federal Level**
  Executive Branch (President, Secretary of Education, Department of Education)
  Congress, including Congressional staff members

- **State Level**
  Executive Branch (governors, state departments of education—testing personnel, curriculum personnel)
  Legislatures

- **Local District Level**
  District administrators and professional staff (superintendents, testing personnel, curriculum personnel)
  School board members

- **Local School Level**
  Principals
  Teachers

- **General Public**
  Groups associated with schools (PTA, advocacy groups)
  Parents
  Taxpayers
  The business and industrial community

- **Members of the Press**
  Newspaper reporting personnel
  Television news personnel
  Radio news personnel

- **Educational Research Personnel**
  Policy analysts
  Psychometricians

As noted earlier, each of these audiences has differing interests and needs for information about NAEP and its results, has differing access to potential dissemination vehicles, and has differing capacity to comprehend and use NAEP information presented in various forms and formats. Audience must, of necessity, be a major dimension in any research on NAEP reporting and dissemination. It also must be realized that these audiences vary in their homogeneity, so that generalizations concerning interests in and needs for NAEP information will be more appropriate for some of these audiences than for others.

Strategies for Research on Reporting and Dissemination of NAEP Results

Research is one area of endeavor where form follows function. Particular strategies obviously will be appropriate in the pursuit of some research questions and of little value in the pursuit of others. Nonetheless, it may be helpful to list some possibilities before proposing that they be pursued in conjunction with particular inquiries. In doing so, it will be immediately clear that some strategies have been applied repeatedly in previous studies on NAEP reporting while others have not. No claim is made that this listing is exhaustive:

- User surveys
  Mail
  Telephone
- Focus group discussions
- Think-aloud interviews
- Large group meetings (e.g., with state assessment directors or with media representatives)
- Content Analyses (e.g., of press reports, Board of Education meeting minutes, etc.)
- Simulations (e.g., of development of press reports in response to press releases)

The Structure of a Research Program

A program of research on NAEP reporting and dissemination could be structured in terms of the intersection of the three dimensions described in the section above. Considering an audience for NAEP results, the research questions to be pursued would be selected from those enumerated earlier, and to pursue those questions, one or more research strategies would be applied. Many more combinations of these dimensions would be feasible than will be of interest, and many more will likely be of interest than can be pursued in a research program with limited resources. In consideration of budget

constraints not only on a program of research but on potential NAEP reporting strategies, selection among possibilities therefore will be critical. Obviously, some studies will be of greater importance than others and will, therefore, gain priority.

Tables 1 through 9 enumerate a set of potential studies on reporting and disseminating NAEP's results for each of NAEP's major audiences, by research question and research strategy. This listing is intended to suggest what might be done, and is presented without regard to priority, and with no consideration of cost. It is sometimes useful to enumerate, without constraint, what might be done, and then impose restrictions and set priorities after studying the range of possibilities. These tables have been constructed in that spirit.

The proposed research strategies reflect the author's beliefs concerning approaches that should be useful and might be feasible. The table entries are, admittedly, speculative concerning the likelihood that any particular study would be administratively feasible, would produce rates of response that would support interpretation of results within acceptable levels of bias error, and would produce findings that contribute to greater understanding of how best to report and disseminate NAEP's results.

Narrative descriptions of some suggested studies that would seem to be of particular interest are presented in a penultimate section of this report. Here again, suggested priorities are a matter of judgment. Judgments were made with due consideration of the findings of previous research on NAEP dissemination and reporting that have queried major users of NAEP's results and have produced some information abut the success of current NAEP reports in stimulating accurate interpretations.

Table 1. Audience: Federal Executive Branch

Research Strategy/Question	User Surveys Mail	User Surveys Phone	Focus Groups	Think-Aloud Interviews	Large Group Meetings	Content Analyses	Simulations
What to Report?							
Interest		x	x				
Utility		x					
Understandable				x			
Validly Interpret				x			
Preference		x	x				
How to Report?							
Interest		x					
Utility		x					
Understandable				x			
Validly Interpret				x			
Preference		x		x			
Dissemination Vehicle							
Accessibility	x						
Regular Use	x						
Preference	x		x				
Feasibility			x				

Table 2. Audience: Congressional Staff Members

Research Strategy/Question	User Surveys Mail	User Surveys Phone	Focus Groups	Think-Aloud Interviews	Large Group Meetings	Content Analyses	Simulations
What to Report?							
Interest		x	x				
Utility		x					
Understandable				x			
Validly Interpret				x			
Preference		x	x				
How to Report?							
Interest		x					
Utility		x					
Understandable				x			
Validly Interpret				x			
Preference		x		x			
Dissemination Vehicle							
Accessibility	x				x		
Regular Use	x				x		
Preference	x				x		
Feasibility				x			

Table 3. Audience: State Executive Branch

Research Strategy/Question	User Surveys Mail	User Surveys Phone	Focus Groups	Think-Aloud Interviews	Large Group Meetings	Content Analyses	Simulations
What to Report?							
Interest		x	x		x		
Utility		x	x			x	
Understandable				x			
Validly Interpret				x			
Preference		x	x		x		
How to Report?							
Interest		x		x			
Utility		x					
Understandable				x			
Validly Interpret				x			
Preference		x		x	x		
Dissemination Vehicle							
Accessibility	x				x		
Regular Use	x				x		
Preference	x				x		
Feasibility				x			

Table 4. Audience: State Legislatures

Research Strategy/Question	User Surveys Mail	User Surveys Phone	Focus Groups	Think-Aloud Interviews	Large Group Meetings	Content Analyses	Simulations
What to Report?							
Interest		x					
Utility		x				x	
Understandable							
Validly Interpret							
Preference		x					
How to Report?							
Interest		x					
Utility							
Understandable		x					
Validly Interpret		x					
Preference		x					
Dissemination Vehicle							
Accessibility		x					
Regular Use		x					
Preference		x					
Feasibility		x					

Table 5. Audience: District-Level Administrators and Professional Staff

Research Strategy/Question	User Surveys Mail	User Surveys Phone	Focus Groups	Think-Aloud Interviews	Large Group Meetings	Content Analyses	Simulations
What to Report?							
Interest	x	x					
Utility	x	x		x		x	
Understandable				x			x
Validly Interpret				x			x
Preference		x					
How to Report?							
Interest		x					
Utility		x					
Understandable				x			x
Validly Interpret				x			x
Preference		x		x			
Dissemination Vehicle							
Accessibility		x					
Regular Use		x					
Preference		x					
Feasibility		x					

Table 6. Audience: School Principals and Teachers

Research Strategy/ Question	User Surveys Mail	User Surveys Phone	Focus Groups	Think-Aloud Interviews	Large Group Meetings	Content Analyses	Simulations
What to Report?							
Interest		x (Pr)	x				
Utility				x			
Understandable				x			x
Validly Interpret				x			x
Preference		x (Pr)	x				
How to Report?							
Interest			x				
Utility			x				
Understandable				x			x
Validly Interpret				x			x
Preference			x				
Dissemination Vehicle							
Accessibility		x (Pr)	x (Teach)				
Regular Use		x (Pr)	x (Teach)				
Preference		x (Pr)	x (Teach)				
Feasibility		x (Pr)	x (Teach)				

Table 7. Audience: General Public

Research Strategy/Question	User Surveys Mail	User Surveys Phone	Focus Groups	Think-Aloud Interviews	Large Group Meetings	Content Analyses	Simulations
What to Report?							
Interest		x	x	x			
Utility							
Understandable			x	x			x
Validly Interpret				x			x
Preference		x	x	x			
How to Report?							
Interest		x	x	x			
Utility							
Understandable			x	x			x
Validly Interpret				x			x
Preference		x	x	x			
Dissemination Vehicle							
Accessibility		x					
Regular Use		x					
Preference		x					
Feasibility							

Table 8. Audience: Members of the Press

Research Strategy/Question	User Surveys Mail	User Surveys Phone	Focus Groups	Think-Aloud Interviews	Large Group Meetings	Content Analyses	Simulations
What to Report?							
Interest			x			x	
Utility			x			x	x
Understandable				x			
Validly Interpret				x			
Preference			x				x
How to Report?							
Interest				x		x	x
Utility						x	x
Understandable				x			
Validly Interpret				x			
Preference				x		x	x
Dissemination Vehicle							
Accessibility			x				
Regular Use			x				
Preference			x				
Feasibility			x				

Table 9. Audience: Education Research Personnel

Research Strategy/Question	User Surveys Mail	User Surveys Phone	Focus Groups	Think-Aloud Interviews	Large Group Meetings	Content Analyses	Simulations
What to Report?							
Interest		x					
Utility		x					
Understandable			x				
Validly Interpret			x				
Preference		x				x	
How to Report?							
Interest		x					
Utility		x					
Understandable			x				
Validly Interpret			x				
Preference		x				x	
Dissemination Vehicle							
Accessibility		x					
Regular Use		x					
Preference		x					
Feasibility		x					

Some Recommended Studies

By law and tradition, NAEP serves a variety of purposes. Its original mandate was to inform the public about the status and progress of what young people in this nation know and understand. NAEP was to be a social indicator of the knowledge of this nation's youth, a role still implied by its sub-appellation "The Nation's Report Card." NAEP's implications concerning the health of the nation's schools were to be indirect, since it was to assess not only what students learned in school but types of knowledge and skills they might gain through their experience in the larger society as well (Jones 1997). More recently, NAEP has gained more direct policy relevance for the nation's public schools as it has assumed the role of a major stimulant for, and indicator of, the progress of school reform.

The development of the state-based component of NAEP in the 1990s is consistent with the increasing policy relevance of NAEP results. The consequent growth of state-level personnel in governors' offices, state legislatures, and state departments of education as major audiences for NAEP is therefore to be expected. It is not surprising that many of the studies on reporting and dissemination of NAEP results conducted in the past have focused principally on state education personnel as an audience.

In recommending some high-priority studies on NAEP reporting and dissemination, this report posits three principal objectives for NAEP and two associated mechanisms whereby NAEP results are disseminated. First, as already noted, NAEP serves a public reporting function. It is therefore important to learn how NAEP results are communicated to the public, and through pursuit of the research questions enumerated earlier, about the effectiveness of that reporting. Second, NAEP has the potential to serve an instructional policy function—influencing curricular and instructional choices in schools that are made by principals and teachers. It is important to understand how such persons learn about NAEP results and the effectiveness of NAEP reporting and dissemination in providing the kinds of information that could affect their decisions. Third, NAEP has the potential to influence education policy at state and national levels, through the federal executive branch, the Congress, the executive offices of state governments, and state legislatures. Again, it is essential to understand how personnel in these bodies learn about NAEP results and about the effectiveness of NAEP in providing the kinds of information that persons in such bodies can understand and use.

The mechanisms for NAEP reporting and dissemination that might reach these diverse audiences consist not only of the direct reporting strategies that are conducted by NCES and NAGB, but of indirect dissemination through the public media as well. Both must be investigated.

Recognizing that far more could usefully be learned about NAEP reporting and dissemination than current resources will support, what follows is a proposed series of studies listed in order of decreasing priority. This series of studies will *not* address the interests of all of the audiences that are delineated in tables 1 through 9.

Based on the results of past studies on errors and inaccuracies in press reports of NAEP results, the highest priority should be placed on learning how to improve the quality of such reports. Here, the term "the press" includes all public media, whether print, television, or radio.

Research on the influence of press releases would be pursued through detailed content analysis of such reports and of subsequent news stories on NAEP in the print press, on television, and on radio. Of interest would be such issues as the frequency with which press release statements were reported verbatim, the frequency with which specific data displays were reported verbatim, the degree to which public reports on NAEP went beyond the content of press releases, and the influence of informal statements by federal government officials and NAGB members on public reporting of NAEP results.

Second in priority, should be gaining understanding of ways to make NAEP results more understandable and useful to school personnel, particularly principals and teachers, so that NAEP's findings concerning the strengths and weaknesses of students' subject-matter knowledge and skills can more directly be applied in assessments at local levels and, ultimately, in improving curriculum and instruction. This proposal is grounded in the recognition that NAEP's content framework is based on a national consensus concerning what students at particular grade levels should know and be able to do in various subject areas, but that NAEP's results are reported at levels of aggregation that are unlikely to be of interest or value to teachers and principals. Nonetheless, if teachers and principals can learn more about the content that NAEP assesses and the ways in which students' knowledge and skills are assessed, they might apply that knowledge in structuring classroom-based or school-based local assessments, and might thereby learn how their own students perform on items and exercises that reflect NAEP's national consensus on valued educational goals.

Third in priority would be gaining information from citizens, parents, and members of the business community on the content and format of NAEP results they find to be comprehensible, of interest, and validly interpretable. Were this information to be available, it could be used to structure the reporting and dissemination of NAEP findings in ways that would enhance the effectiveness of media-based reporting, and of more direct dissemination strategies, including the World Wide Web. Although it is likely impractical to consider the direct public dissemination of NAEP findings to citizens and parents through print reports, as more households gain internet access, it is increasingly practical to consider direct public dissemination through electronic means, such as the World Wide Web.

Fourth in priority, would be additional study of reporting and dissemination of NAEP results to state education personnel. This audience is particularly important for several reasons. First, the state assessment component of NAEP provides many states with a comparative national reference on the achievement status of the state's students in selected subjects and at particular grade levels. Second, the growth of state-wide accountability programs, coupled with the adoption of uniform statewide curricula, has

increased the policy relevance of state NAEP results. Many curricular decisions are now made at state, rather than local, levels.

The lowest priority to this study only because the greatest amount of research on NAEP reporting to date has focused on state education personnel. In addition, currently ongoing research on NAEP reporting (an NCES-supported study in which Howard Wainer and Ronald Hambleton are co-principal investigators) is addressing this group.

A synopsis of proposed studies follows:

Priority 1: Research on Reporting Through Public Media. It is likely that members of the general public and many policymakers receive their information on NAEP results either principally or solely through the public media. Indeed, John F. Jennings, former Chief Counsel to the House Education and Workforce Committee, during a symposium at the 1996 annual meeting of the American Educational Research Association, noted that members of Congress received more information about NAEP results through the *Washington Post* than through any other vehicle, including reports sent by NCES. Unfortunately, as noted earlier, studies of press reports on NAEP have revealed rampant inaccuracies and frequent unwarranted inferences (Koretz and Deibert 1993; Jaeger 1996). It is therefore critical for NCES and NAGB staff to learn how they can improve the quality and accuracy of press reports on the findings of the National Assessment.

Since the press will be naturally suspicious of any activities of a government agency that are designed to influence reporting, research on this topic will have to be designed carefully and cautiously. In fact, it might be necessary to divorce the federal government completely from the research by having it designed, conducted, and reported by an independent agent or agency. Hambleton and Slater (1995) indicated that their attempts to secure the cooperation of members of the press in a federally-supported study on NAEP reporting were singularly unsuccessful. They reported: "…several newspaper writers who we did contact declined our invitations to participate. They said they preferred asking questions to answering them and would not participate in the study" (4). A fruitful alternative to government sponsorship of inquiry in this area might be foundation-sponsored research. In conducting preliminary inquiries with members of the print press on a potential study on reporting of NAEP results, a colleague and I learned that newspaper reporters are wary of working on press reports with personnel from the agencies that are the subject of their reporting. Journalistic canons require reporters to keep an arms-length distance from the objects of their journalism.

Among the research questions to be pursued are the following:

- To what degree does the content of press releases by NCES or NAGB influence the content of subsequent news reports by the press, television news and radio news?

- How is the interest of media personnel in NAEP findings influenced by the kinds of NAEP results that are reported and the format in which results are reported?

- What are the preferences of members of the media among various forms of NAEP results and formats for reporting NAEP results?

- Among potential vehicles for reporting NAEP results, what are most accessible to media personnel, what vehicles do they regularly use, what do they prefer, and which are most feasible for reporting particular kinds of NAEP results?

- How is the ability of media personnel to understand and validly interpret NAEP results influenced by various forms of reporting and formats for reporting the results?

- What forms of reporting and formats for reporting NAEP results do media personnel find to be most useful?

Four research strategies are suggested for pursuing these questions (see table 8). Since it is doubtful that members of the press and other media personnel would engage in either mail or telephone interview surveys on the topics of these research questions, these strategies are not recommended. Instead, it is proposed to hold small conferences of the sort that have been successfully supported by the Ford Foundation on education of the press to engage media personnel in focus groups, think-aloud interviews, and simulations during which they would compose simulated stories on NAEP results. The focus groups would be used to obtain factual information on the accessibility and use of and preferences among various dissemination vehicles for NAEP results, and to obtain judgments concerning respondents' interest in and the utility of various of NAEP information and formats for reporting NAEP results. The think-aloud interviews would be used to obtain information on the influence of form and format for reporting NAEP results on participants' abilities to understand the reported information and to interpret it validly. A fourth research strategy would involve content analyses of press reports and resulting news stories in an attempt to infer the influence of the former on the latter.

Priority 2: Making NAEP Reporting More Understandable and Useful to School Curriculum and Instruction Personnel. To date, the majority of research on NAEP reporting has focused on state-level assessment personnel and curriculum specialists (DeVito 1997; Hawkins 1995; Hambleton and Slater 1995). With the exception of a small survey and focus-group investigation conducted by the Widmeyer Group in conjunction with a NAEP marketing study (Widmeyer 1993), the possibility that NAEP results either have been or could be reported in ways that influence school curriculum and instruction more directly, through school principals and teachers, has

largely been overlooked. This study would remedy that deficiency by seeking to discover how NAEP results can be reported so as to be understandable and useful to school personnel and how reporting of NAEP results might enhance its influence on school curricula and instruction.

The research questions that would be addressed through this study include:

- What is the degree of familiarity of school teachers and principals with NAEP and its results?
- What is the relationship, if any, between the form and format of NAEP reporting and the interest of teachers and principals in its results?
- What are the preferences of teachers and principals among alternative forms and formats of NAEP reporting?
- How useful do teachers and principals find NAEP results, and how might that utility be influenced by alternative forms of reporting and reporting formats?
- What is the relationship, if any, between the form and format of NAEP reporting and the ability of teachers and principals to understand NAEP results and to validly interpret them?
- What dissemination vehicles for NAEP results are accessible to teachers and principals, regularly used by teachers and principals, and preferred by teachers and principals?
- Through what dissemination vehicles would it be feasible to disseminate various kinds of NAEP results to teachers and principals?

The research strategies that would be used in this study would include telephone interviews of school principals; focus-group discussions with principals and separate focus-group discussions with teachers; personal, think-aloud interviews of principals and, separately, of teachers; and finally, studies in which principals and teachers would be asked to interpret and draw conclusions from simulated NAEP reports that incorporated alternative reporting forms and formats. The decision to work with teachers and principals separately is influenced by consideration of the status hierarchy in public schools, which might intimidate or otherwise bias the responses of teachers in a joint setting with principals. The proposed application of these strategies to the research questions listed above is summarized in table 6.

Priority 3: Reporting to the Public. To determine how well NAEP is succeeding in achieving its public information function, a study that focuses on the general public, with parents of school children as a subpopulation might be useful. Virtually all of the research questions enumerated earlier—concerned with what to report, how to report, and

dissemination vehicles—should be pursued. The four recommended research strategies are a telephone interview survey, a number of focus groups, a series of individual, face-to-face, think-aloud interviews, and a study to gauge interpretations of simulated NAEP reports that made use of alternative reporting forms and formats. A subset of persons interviewed by phone would be recruited for focus-group research and for personal, think-aloud interviews. Cluster sampling of interviewees would be necessary for this second phase of the study. Jaeger et al. (1993) successfully obtained information from parents of school children concerning their desires for information about the schools their children attended and their abilities to correctly interpret reports on the quality of schools. These researchers used telephone interviews followed by face-to-face interviews during which sampled parents described their interpretations of simulated school reports. A similar strategy, augmented by several focus-group sessions, is proposed for the study of NAEP reporting and dissemination. Sampled members of the public would be sent letters indicating that they would be called to engage in interviews concerning the achievement of the nation's students and would receive a nominal stipend for their time. Those who agreed to be interviewed would be sent samples of real and simulated NAEP summary reports as well as real and simulated NAEP press reports on results. Telephone interviews would seek information on:

- Respondents' interest in NAEP results and on the degree to which that interest was affected by choices of what to report and the format in which results were reported;
- Respondents' preferences among alternative forms of reporting and among alternative formats for reporting of NAEP results;
- Respondents' access to various potential NAEP reporting vehicles;
- The regularity of respondents' use of various potential NAEP reporting vehicles; and
- Respondents' preferences among various potential NAEP reporting vehicles.

In the second phase of this study, cluster samples of parents and members of the general public would be asked to participate in a focus group or would be interviewed in person and asked to "think aloud" while they read and interpreted various real and simulated reports on NAEP results. This phase of the study also would make use of real and simulated excerpts from NAEP summary reports and of real and simulated press reports on NAEP results. Respondents would be asked to (1) describe and interpret what they read, and (2) to describe the implications of what they read by drawing conclusions about student achievement, the comparative achievement of various groups, concomitants of student achievement, and the quality of the schools. Relationships between forms and formats of reports and respondents' abilities to comprehend reported

information and to make valid interpretations of reported information would be investigated.

One type of report that would be included in this research would be a focused report that included multiple examples of student work in response to NAEP exercises. These would be selected and organized to help readers answer such questions as: "Do students know basic facts?" and "Can students solve word problems?". John Dossey, a co-author of many NAEP reports, has noted that the broad item sampling used in the design of NAEP assessments would support the production of such reports and has suggested that such focused reports are likely to be of substantial interest to parents and other lay readers.

Priority 4: Further Research with State Education Personnel. Although some research on NAEP reporting has focused on state education personnel (DeVito 1997; Hambleton and Slater 1995; Hawkins 1995) the research to date has been limited in several ways. For instance, subjects have been selected on the basis of convenience rather than through scientific sampling and, with the exception of the Hambleton and Slater study, the research has focused largely on respondents' preferences among several forms of reporting NAEP results rather than their abilities to correctly interpret results or draw appropriate inferences from NAEP reports. The Hambleton and Slater research clearly revealed the existence of a problem since so many of their respondents incorrectly interpreted information in the NAEP report that was shown to them. However, their research did not provide evidence on the likely effectiveness of alternatives. More must be learned about how NAEP reporting forms and formats can affect interpretability and stimulate or hinder the formation of appropriate inferences.

The research questions that would be addressed in this study include the following:

- Among alternative forms of reporting NAEP results, which stimulate greatest interest, are found to be most useful, are most frequently found to be understandable, most frequently stimulate correct interpretation, and are most preferred by state education personnel?

- The same questions would be asked of alternative formats for reporting NAEP results.

- Among potential dissemination vehicles for NAEP results, which are most accessible to state education personnel, which do state education personnel regularly use, and which are most preferred?

As indicated in table 3, a variety of research strategies would be used in this study. Questions concerning interest engendered by various NAEP reporting formats and forms, as well as preferences among alternatives could be investigated through telephone surveys, focus groups convened at professional meetings, and in general sessions of such meetings. The question of the utility of alternative forms and formats of NAEP reporting has two components. One component concerns perceived utility and the other actual

evidence of usefulness. Telephone surveys and focus groups would be appropriate strategies for collecting information on perceived utility. Evidence that NAEP results in various formats were factually found to be useful by state education personnel could be obtained through content analyses of state reports on student achievement, content analyses of state board of education meeting minutes, and content analyses of state assessments and education regulations related to student assessment. Information on the dissemination vehicles for NAEP results that are accessible to, used by, and preferred by state education personnel could be obtained through telephone interviews or in general sessions at professional meetings of such personnel. As in studies described earlier in this report, think-aloud interviews would be used to investigate the effects of reporting form and format on the perceived comprehensibility of NAEP results as well as the validity of participants' interpretation of those results.

In this study, as in several described earlier, stimulus materials would include not only current NAEP reports, but simulated reports that varied systematically in the form of information reported as well as in reporting format. Materials would be sent to respondents in advance of telephone interviews, and they would be asked to study the materials in preparation for the interviews. Since past studies on NAEP reporting involving state education personnel have realized good rates of cooperation, it is doubtful that subjects in this study would have to be compensated for their participation.

Concluding Remarks

Valid use of any assessment demands effective communication of its results and accurate interpretation of its findings. Although the agencies responsible for NAEP have worked with great effectiveness to ensure that NAEP's content frame is consistent with the latest conceptions of appropriate subject-matter curricula, that NAEP's exercises are closely linked to its content frame, that students in federally-protected groups are not unduly disadvantaged by the form and format of NAEP's exercises, and that NAEP's technical measurement properties are sound, have contributed to these agencies being less successful in their ability to ensure that NAEP results are presented in ways that are clearly understood and correctly interpreted by its constituencies. The research reviewed in this paper suggests that many NAEP audiences find NAEP results difficult to comprehend and frequently err in their interpretations of its findings. Although NCES has sponsored some research on the effectiveness of NAEP reporting, more is now known about the magnitude of the problem than about its solution. Current suggestions for improvements to NAEP reporting appear to be reasonable but are largely speculative. A research program of the sort proposed here is needed to learn what to report, how to report it, and how to disseminate what is reported. Only by confirming that NAEP's audiences can comprehend its results will we be certain that valid interpretations are possible.

References

Adler, J. 1990, Fall/Winter. Creating problems. *Newsweek Special Issue: How to Teach Our Kids*. 16–18, 22.

Angoff, W. H. 1971. Scales, norms, and equivalent scores. In R. L. Thorndike (ed.), *Educational measurement* (2nd. ed.). Washington, DC: American Council on Education.

Aschbacher, P. R. & Herman, J. L. 1991. *Guidelines for effective score reporting*. CSE Technical Report 326. UCLA Center for Research on Evaluation, Standards, and Student Testing.

Beaton, A. E. & Johnson, E. G. 1992. Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement, 29,* 163–175.

Beaton, A. E. & Allen, N. L. 1992. Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17,* 191–204.

Bock, R. D. 1997. Domain scores: A concept for reporting the National Assessment of Educational Progress results. In Glaser, R., Linn, R., & Bohrnstedt, G. (eds.), *Assessment in Transition: Monitoring the Nation's Educational Progress, Background Studies*. Stanford, CA: National Academy of Education, 81–102.

Bock, R. D., Thissen, D., and Zimowski, M. F. 1997. IRT estimation of domain scores. *Journal of Educational Measurement, 34,* 197–211.

Bourque, M. L. and Garrison, H. H. 1991. *The levels of mathematics achievement: Initial performance standards for the 1990 NAEP mathematics assessment*. Washington, DC: National Assessment Governing Board.

Chang, L. 1996. Does a standard reflect minimal competency of examinees or judge competency? *Applied Measurement in Education, 9,* 161–173.

Cizek, G. J. 1993. *Reactions to National Academy of Education Report, "Setting Performance Standards for Student Achievement."* Unpublished Manuscript.

DeMauro, G. E. 1995, March. Construct validation of minimum competence in standard setting. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

DeVito, P. J. 1997. The future of the National Assessment of Educational Progress from the States' Perspective. In Glaser, R., Linn, R., & Bohrnstedt, G. (eds.). *Assessment in transition: Monitoring the Nation's Educational Progress, Background Studies*. Stanford, CA: National Academy of Education, 31–46.

Ebel, R. L. 1962. Content standard test scores. *Educational and Psychological Measurement, 22*, 15–25.

Education Commission of the States 1970. *National Assessment of Educational Progress, Report 1, 1969–1970 Science: National Results and Illustrations of Group Comparisons*. Denver, CO: Author.

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

Forsyth, R., Hambleton, R. K., Linn, R. L., Mislevy, R., and Yen, W. 1996. *Design feasibility team: Report to the National Assessment Governing Board*. Washington, DC: National Assessment Governing Board.

Haertel, E. 1997, August. Unpublished memorandum to the National Assessment Governing Board staff and selected board members dated 6 August, 1997. Stanford, CA: Stanford University, School of Education.

Hambleton, R. K. and Slater, S. C. 1995. *Are NAEP executive summary reports understandable to policymakers and educators?* Amherst, MA: Project report under National Center for Education Statistics Contract No. RS90159001.

Hawkins, E. 1995. Impact of the 1992 National Assessment of Educational Progress trial state assessment. In Glaser, R. Linn. R. L., & Bohrnstedt, G. (eds.), *Quality and Utility: The 1994 Trial State Assessment in Reading, Background Studies*. Stanford, CA: National Academy of Education, 403–427.

Impara, J. C., and Plake, B. S. 1996, April. Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Jaeger, R. M. 1989. Certification of student competence. In R. L. Linn (ed.), *Educational Measurement* (3rd. ed., 485–514). New York: American Council on Education and Macmillan.

Jaeger, R. M. 1996, April. Reporting large-scale assessment results for public consumption: Some propositions and palliatives. Presented at the 1996 annual meeting of the National Council on Measurement in Education, New York.

Jaeger, R. M., Gorney, B., Johnson, R. L., Putnam, S. E., and Williamson, G. 1993. *Designing and Developing Effective School Report Cards: A Research Synthesis*. Kalamazoo, MI: Center for Research on Education Accountability and Teacher Evaluation, Western Michigan University.

Jones, L. 1997. The National Assessment of Educational Progress, origins and prospects. In Glaser, R. Linn. R. L., & Bohrnstedt, G. (eds.), *Assessment in Transition: Monitoring the Nation's Educational Progress, Background Studies*. Stanford, CA: National Academy of Education, 1–18.

Kane, M. 1993. *Comments on the NAE Evaluation of the NAGB Achievement Levels*. Unpublished Manuscript.

Kane, M. 1994. Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–461.

Kirsch, I. S. and Jungeblut, A. 1986. *Literacy: Profiles of America's Young Adults*. Princeton, NJ: Educational Testing Service.

Koretz, D. M. and Deibert, E. 1993. *Interpretations of National Assessment of Educational Progress (NAEP) Anchor Points and Achievement Levels by the Print Media in 1991*. Santa Monica, CA: RAND.

Linn, R. L. 1996. Validating inferences from NAEP achievement level reporting. Presented at a National Research Council Workshop on NAEP Achievement Levels: Setting Consensus Goals for Academic Achievement. Washington, DC: December 6 and 7, 1996.

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

Martinez, M. E. and Mead, N. A. 1988. *Computer competence: The first national assessment*. Princeton, NJ: Educational Testing Service, Report No. 17-CC-01.

Messick, S., Beaton, A. E., and Lord, F. 1983. *A New Design for a New Era*. Princeton, NJ: Educational Testing Service.

Mislevy, R. J. 1996. Implications of market-basket reporting for achievement level setting. Paper presented at a workshop on Setting Consensus Goals for Academic Achievement, under the sponsorship of the National Research Council Committee on the Evaluation of NAEP, Washington, DC, December 6–7, 1996.

Mislevy, R. J., Johnson, E. G., and Muraki, E. 1992. Scaling procedures in NAEP. *Journal of Education Statistics, 17,* 131–154.

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. 1992. Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133–161.

Mullis, I. V. S., Oldefeldt, S. J., and Phillips, D. L. 1977. *What students know and can do: Profiles of three age groups.* Denver, CO: National Assessment of Educational Progress.

Mullis, I. V. S., Dossey, J. A., Owen, E. H. and Phillips, G. W. 1993. *The 1992 Mathematics Report Card.* Washington, DC: U. S. Department of Education.

National Assessment of Educational Progress 1978. *Three National Assessments for Science: Changes in Achievement 1969–1977.* Denver, CO: Author.

National Assessment Governing Board 1997. *1996 Science Performance Standards: Achievement Results for the Nation and the States.* Washington, DC: Author.

National Assessment Governing Board 1997. *What Do Students Know? 1996 NAEP Science Results for 4th, 8th, & 12th Graders.* Washington, DC: Author.

National Center for Education Statistics 1997. *The NAEP Guide, 1997 Edition.* NCES-97-990, Calderone, J., King, L. M., and Horkay, N. (eds.). Washington, DC.

Phillips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H. and Barton, P. E. 1993. *Interpreting NAEP Scales.* Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Public Law 100-297 1988. *National Assessment of Educational Progress Improvement Act* (Article No. USC 1221). Washington, DC.

Quereshi, M. Y., and Fisher, T. L. 1977. Logical versus empirical estimates of item difficulty. *Educational and Psychological Measurement, 37,* 91–100.

Reese, C. M., Miller, K. E., Mazzeo, J., and Dossey, J. A. 1997. *NAEP 1996 Mathematics Report Card for the Nation and the States.* Washington, DC: National Center for Education Statistics.

Rothman, R. 1991, October. Four-fifths of students fail to attain proficiency in math, report shows. *Education Week,* 14–15.

Shanker, A. 1990. The end of the traditional model of schooling and a proposal for using incentives to restructure our public schools. *Phi Delta Kappan, 71* (5), 344–348.

Shepard, L. A., Glaser, R., Linn, R. L., and Bohrnstedt, G. 1993. *Setting Performance Standards for Student Achievement: An Evaluation of the 1992 Achievement Levels*. A report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment. Stanford, CA: Stanford, California, National Academy of Education.

Stufflebeam, D. L., Jaeger, R. M., and 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, MI: Western Michigan University.

Taube, K. T. and Newman, L. S. 1996, April. *The Accuracy and Use of Item Difficulty Calibrations Estimated from Judges' Ratings of Item Difficulty.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Thorndike, R. L. 1980. Item and Score Conversion by Pooled Judgment. *Proceedings of the Educational Testing Service Conference on Test Equating*, Princeton, NJ.

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

U. S. General Accounting Office 1993. *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations*. (GAO/PEMD–93–12). Washington, D C: Author.

Wainer, H. 1997. Improving tabular display, with NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics, 22*, 1–30.

Ward, B. 1980. *Major Information Needs of National Assessment Audiences and Ways to Enhance the Assessment's Utility in Meeting those Needs*. Denver, CO: Education Commission of the States.

Wheeler, P. 1991, April. *The Relationship Between Modified Angoff Knowledge Estimation Judgments and Item Difficulty Values for Seven NTE Specialty Area Tests*. Paper presented at the Annual Meeting of the California Educational Research Association, San Diego, CA. ERIC Document Reproduction Service No. ED 340 745.

Widmeyer Group, Inc. 1993. *Dissemination Strategies for the National Assessment of Educational Progress*. Report prepared by the Widmeyer Group, Inc. for the National Assessment Governing Board. Washington, DC: Author.