THE NATION'S REPORT CARD

NAEP

National Assessment Governing Board

Student Performance Standards on the National Assessment of Educational Progress:

Affirmation and Improvements

Washington, DC • November 2000

A Study Initiated To Examine a Decade of Achievement Level Setting on NAEP

What Is The Nation's Report Card?

The Nation's Report Card, the National Assessment of Educational Progress (NAEP), is the only nationally representative and continuing assessment of what America's students know and can do in various subjects. Since 1969, assessments have been conducted periodically in reading, mathematics, science, writing, history/geography, and other fields. By making objective information on student performance available to policymakers at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement is collected under this program. NAEP guarantees the privacy of individual students and their families.

NAEP is a congressionally mandated project of the National Center for Education Statistics, the U.S. Department of Education. The Commissioner of Education Statistics is responsible, by law, for carrying out the NAEP project through competitive awards to qualified organizations. NAEP reports directly to the Commissioner, who is also responsible for providing continuing reviews, including validation studies and solicitation of public comment, on NAEP's conduct and usefulness.

The National Assessment Governing Board (NAGB) was established under section 412 of the National Education Statistics Act of 1994 (Title IV of the Improving America's Schools Act of 1994, P.L. 103–382). The Board was established to formulate policy guidelines for NAEP. The Board is responsible for selecting subject areas to be assessed, developing assessment objectives, identifying appropriate achievement goals for each grade level and subject tested, and establishing standards and procedures for interstate and national comparisons.

The National Assessment Governing Board

Student Performance Standards on the National Assessment of Educational Progress:

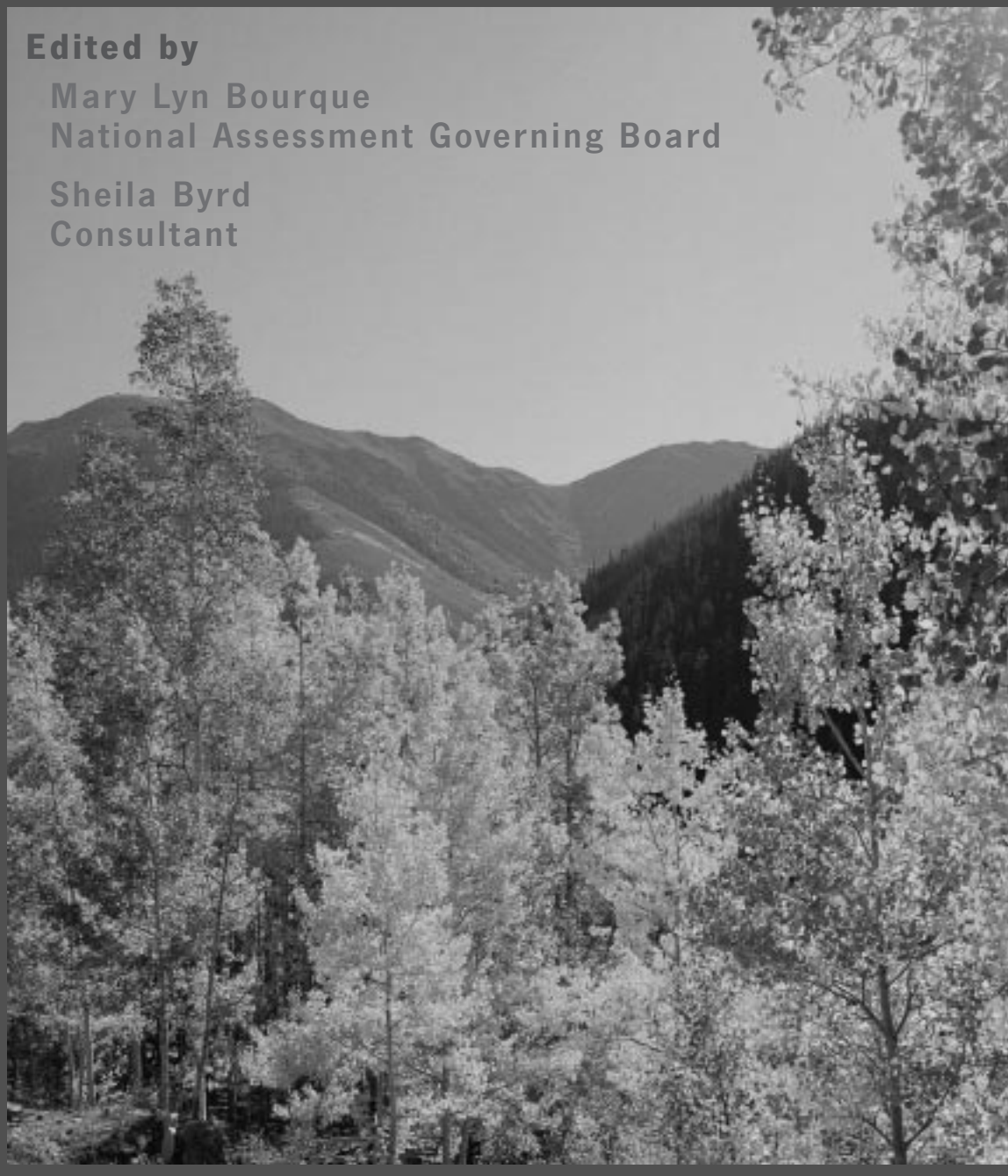
Affirmation and Improvements

A Study Initiated To Examine a Decade of Achievement Level Setting on NAEP

Edited by

Mary Lyn Bourque
National Assessment Governing Board

Sheila Byrd
Consultant



Washington, DC • November 2000

November 2000

Suggested Citation

Bourque, M.L. and Byrd, S. (Eds.)
Student Performance Standards on the National Assessment of Educational Progress: Affirmations and Improvements.
Washington, DC: National Assessment Governing Board, 2000.

For more information contact:

National Assessment Governing Board
202–357–6938

For ordering information on this report, write:

National Assessment Governing Board
800 North Capitol Street, NW
Suite 825
Washington, DC 20002–4233

This report is also available on the World Wide Web.
http://www.nagb.org

Table of Contents

Introduction

Over the past decade, the National Assessment Governing Board (NAGB) has used the most prevalent models for setting standards on the National Assessment of Educational Progress (NAEP). During that time, NAGB has made many refinements to the original process and improved that process considerably. However, experts continue to differ over whether alternative models would be better and/or would accomplish the Board's policy goals for NAEP any more effectively.

There is no current recommendation to abandon the existing model in the hope of finding the *perfect model.* Nevertheless, the Achievement Levels Committee and the Board decided to take a deliberate look at their progress 10 years after setting standards on NAEP. During that decade, the Board has set standards in seven subjects: reading, mathematics, U.S. history, world geography, science, civics, and writing.

The articles in this report were commissioned by the Board based on the Achievement Levels Committee's recommendations. The articles present possible avenues of exploration that could result in fruitful and productive endeavors.

The information-gathering phase did not consist solely of the written papers. The Committee encouraged other forms of data gathering, including examination of other standard-setting activities (perhaps even observation of some in other contexts if possible); advice from policy groups; content analysis of extant documents; focus groups; public commentary; and written technical papers. The Board directed that this effort examine three areas in depth: (1) an integrated view of Board policy on standard setting as a response to the major criticisms, (2) public perception of the standards, and (3) a review of extant models for standard setting.

Section 2, *Reporting NAEP by Achievement Levels: An Analysis of the Policy and the External Reviews,* by William Brown, Brownstar, Inc., synthesizes the meaning of the NAEP achievement levels in Board policies from the inception of the standards in 1990 to the present. It clarifies the meaning of concepts related to the achievement levels and articulates tacit concepts embedded in the Board's policies. In addition, it catalogs, by subject area, all the major criticisms that have been leveled against the standards by various evaluations, as well as policy commentators and others who have written about the levels' shortcomings. Brown's paper reflects the Board's policy development as an ongoing conversation and response to the various evaluations and critiques.

Mark D. Reckase, Michigan State University, has prepared a literature review that examines other models currently available for developing student performance standards. The review in Section 3, *A Survey and Evaluation of Recently Developed Procedures for Setting Standards on Educational Tests,* examines a variety of important questions. Do all models work equally well? Are there tradeoffs with some models that are greater than those with other models? How does one weigh the costs and benefits of one model over another? What models have been researched over the past 10 years that were found not to work so well in the NAEP setting? Is there a theoretical base for any of the newer models? What does the current research say about these models?

Sections 4 to 8 detail public perception and use of the standards. Robert A. Forsyth, University of Iowa, reviewed the standards that commercial test publishers are promoting in their "shelf-tests." Forsyth's paper in Section 4, *A Description of the Standard-Setting Procedures Used by Three Standardized Achievement Test Publishers,* explores a number of relevant questions. Are those standards similar to the NAEP achievement levels? Do they use a similar process for developing the levels? Do the commercial test publishers engage similar panels in developing the standards? Are their results similar to the NAEP results?

Jeffrey M. Nellhaus, Massachusetts Department of Education, reviewed States' activities in setting standards on State assessments in Section 5. His paper, *States With NAEP-Like Performance Standards,* explores which States are using NAEP-like standards and which States have appropriated the actual NAEP standards.

In Section 6, Ronald K. Hambleton and Kevin Meara, University of Massachusetts at Amherst, trace media coverage of the levels over the years. Their paper, *Newspaper Coverage of NAEP Results, 1990 to 1998,* reviewed several hundred press clippings and the press packet materials used by the Government to release the NAEP results. Their work examined questions about what is being reported, whether it is being reported accurately, and whether the reporting enhances the interpretation of NAEP results.

W. James Popham, University of California at Los Angeles, prepared a paper taking a critical look at the reporting mechanisms for NAEP achievement levels. Popham's review in Section 7 pays particular attention to the need for providing clear and accurate information on the large percentages of students performing at the Basic and Below Basic levels. His paper, *Looking at Achievement Levels,* offers creative solutions to this reporting dilemma.

In Section 8, Claudia Simmons and Munira Mwalimu present information they gathered, through focus groups, about the perception of NAEP's publics with respect to the three legislated criteria: reasonable, valid, and informative to the public. Their report, *What NAEP's Publics Have To Say,* systematically reports on the views of State legislative staff; Governors' staff; State assessment personnel; public and private school teachers, administrators, and parents; and the business and industry communities.

The goal of gathering this information was to assist the Board in examining and clarifying its public policy positions and its operational procedures in the area of standard setting for the future. This information, coupled with a study of NAGB's current policies and information from the piloting of alternative models, will be a resource for the Board to use to craft future policy directions for NAEP.

NAEP has a rich and long experience in this area. It is hoped by the Board that the information contained in each of these papers will be of value and utility in setting student performance standards.

SECTION 1

Executive Summary

Sheila Byrd Consultant

November 2000

Executive Summary

As part of its ongoing research and development agenda, the National Assessment Governing Board (NAGB) commissioned the following articles, which explore important aspects of NAGB's achievement levels-setting process, the public's perception of its results, and the ways in which other processes, both State and commercial, compare. The commissioned research also includes an examination of alternative standard-setting methods and their possible applicability for the National Assessment of Educational Progress (NAEP). Finally, modifications to existing achievement levels are considered.

NAGB's policy statement clearly states the Board's understanding of the standard-setting process and the need for continuous evaluation of that process over time:

> The development of achievement levels requires vigilance to ensure that aspects of the level-setting process not be prematurely institutionalized, closing off new ways of thinking about the levels, new ways of expressing assessment frameworks in terms of the levels, new technologies for assessing student performance, interpret-ing NAEP data, and reporting NAEP results.[1]

Although mindful of the value of stability, these articles comprise a thorough evaluation and reconsideration of all aspects of NAGB's standard-setting process, including the Board's policies and practices and their impact on students, education professionals, policymakers, and the public.

The results of the Board's deliberations as this policy review is completed, and the results of any pilot testing of new methods initiated will be used to inform NAGB's decisions on future standard setting, both for NAEP (starting as early as mathematics in 2004) and for the proposed Voluntary National Tests (depending on congressional action and starting no earlier than field testing in 2001).

Following are summaries of each of the commissioned articles, including a summary description of four focus groups that were convened to gather further information about public perception of the standards among primary NAEP audiences.

Review of Governing Board Policies and Practices

Reporting NAEP by Achievement Levels: An Analysis of the Policy and the External Reviews, **William Brown**

This synthesis examines all evaluations of the NAGB achievement levels-setting process conducted since 1990. It provides an integrated view of Board policy on standard setting and the Board's response to major criticisms (organized by model, process, and product). The report suggests that throughout the past decade, NAGB has not simply accepted every change

[1] National Assessment Governing Board (1993). *Developing Student Performance Levels on the NAEP*. (Policy Statement). Washington, DC: National Assessment Governing Board.

suggested by every critic, but rather evaluated each criticism for its potential to improve the process, affirming or changing policies as appropriate.

Brown describes the ways in which the standard-setting process employed by NAGB has evolved since 1992. He cites specific improvements (consistent with the Board's principles for levels setting) in the alignment of frameworks and achievement level descriptions, training of panelists, new forms of feedback, augmentation of the NAEP item pool, better matches of exemplar items and performance levels, and the piloting of other models.

Noting that NAGB made a deliberate choice in favor of qualitative reporting because it would "provide an impetus for change even if the performance levels were not satisfactory initially," Brown concludes by asking the salient policy question (p. 38)

> Clearly the problems that were identified with the initial process were concerns to be addressed, and many of them have been studied. The question to be addressed now is whether the recommendations by critics to abandon the current achievement level-setting process are warranted by the problems identified.

In addition, Brown raises a question that is explored in detail in the research conducted by Mark D. Reckase: Is there a viable and tested model available that will produce more valid and more reliable results?

Public Perception of the Standards

Closely related to the questions about Board policies and practices, two studies examined the ways in which NAEP audiences have perceived the standards throughout their first decade.

What NAEP's Publics Have To Say, Claudia Simmons and Munira Mwalimu, Aspen Systems Corporation

Four focus groups were designed to gather systematic, in-depth information as to whether the achievement levels are reasonable and informative. Claudia Simmons and Munira Mwalimu's summary notes, "Although evidence suggests that the levels are useful and informative . . . the Board felt that it would be helpful to hold several information-gathering sessions around the country with specific NAEP audiences" (p. 185). The focus groups were therefore homogeneous groupings of: (1) Governors' and States' legislative staff (Atlanta); (2) State assessment personnel (Alexandria, Virginia); (3) public and private educators, administrators, and parents (San Francisco); and (4) business leaders and education policymakers (Houston). All four discussion groups focused on the following two topics:

1. The reasonableness of the NAEP achievement levels with regard to three components:
   a. Policy definitions of the achievement levels,
   b. Content descriptions of the achievement levels, and
   c. Relation of NAEP achievement levels to other assessments.

2. Audience experience and reaction to achievement levels with comments focusing on:
   a. Reporting achievement levels, and
   b. Usefulness of achievement levels.

Individual comments are described in detail in Simmons and Mwalimu's summary; general trends are noted here.

With regard to topic 1, State legislative staff, business leaders, and policymakers in general agreed that the policy definitions and content descriptions are reasonable. State assessment personnel and educators/administrators suggested that the policy definitions and content descriptions are laudable goals but also subjective propositions. They support the use of exemplars, but suggested that the exemplars could be improved. There was consensus among the four groups that the NAEP achievement levels cannot be compared with results from other standardized assessments.

Concerning topic 2, there was a strong consensus on the usefulness of the levels, but more varied perceptions of how NAEP data are reported. In Atlanta, the group believed that coverage of NAEP in recent years had increased at the national and State levels, but not much interest had been shown at the local level. There was strong consensus at the Alexandria meeting on the need for NAEP data and the importance of NAEP data to the States. This group was concerned that the media tend to focus only on the "bad news," a concern shared by the San Francisco group. In Alexandria and San Francisco, participants suggested that NAEP results do not reflect variances in State curriculums. All groups supported the release of items.

Newspaper Coverage of NAEP Results, 1990 to 1998, **Ronald K. Hambleton and Kevin Meara, University of Massachusetts at Amherst**

Ten years after the introduction of achievement levels in NAEP score reporting, this paper considers the way in which newspapers have been presenting (and reporters interpreting) NAEP results for the public. Among the questions addressed in this study are:

1. How central are the achievement levels in newspaper reports of NAEP results?
2. What other NAEP information are the newspapers reporting?
3. How well are they reporting it?

Specifically, the research study was designed to answer the following four questions:

1. How have the NAEP press briefing packages changed over the past 10 years?
2. What information has been highlighted in the newspaper accounts of NAEP results?
3. Is there evidence that the NAEP press release materials are being understood and used by the newspapers in their stories?
4. Are the newspapers accurately conveying information about NAEP results to their readers?

More than 500 clippings were reviewed for 16 features: discussion of the standards, reporting of scaled scores, national results, State-by-State information, State-to-national information,

changes over time, curriculum consideration, NAEP over test comparisons, NAEP limitations, multiple-subject reporting, interesting anecdotes or examples, and reporting of gender, race, socioeconomic status, parent, and interaction information.

Hambleton and Meara conclude that the press packages have changed substantially over the years. In general, the study suggests, more information has been included about NAEP itself, the curriculum frameworks, and content of the assessments. Exemplar items have been introduced and the use of graphics in score reporting has increased. In later NAEP releases (1996 and 1998), the press briefing packages have become more informative.

The study confirms that the press reports what is in briefing packages, but also that it misinterprets the findings, often by making unsubstantiated causal inferences. The authors suggest ways that NAGB may point out the difficulties in making such causal inferences from correlational data, acknowledging that the data and reports might be too complex for the public to understand. The authors recommend that NAGB continue to find language and examples to communicate correct interpretations of the levels.

Finally, Hambleton and Meara note problems in explaining the meaning of statistical concepts and scores. The media's confusion over percentiles and cumulative percentages "is passed on to the public." Similarly, the meaning of NAEP scores remains a problem, according to the authors. What is the meaning of a 1- to 3-point change, for example, and how should such a change be interpreted relative to a 5- to 8-point change?

Commercial and State Processes

To augment the scope of the Board's discussions beyond an examination of its own process and therefore stimulate discussion of how its performance standards might be refined in the future, NAGB asked Robert A. Forsyth to examine the processes currently in use by the three major commercial test publishers. Jeffrey M. Nellhaus studied the processes utilized in the States. Analyzing the similarities and differences among the processes will assist NAGB in evaluating the efficacy of its current system and help it determine what, if any, modifications may be necessary. The results of both studies indicate that the NAGB process has had a considerable influence on the methods employed by both the "shelf" and State tests.

A Description of the Standard-Setting Procedures Used by Three Standardized Achievement Test Publishers, Robert A. Forsyth, University of Iowa

Forsyth concludes that the standards in all three cases are similar to NAGB's achievement levels. Two of the three major test publishers use the same number of performance categories as NAGB and similar or identical labels for each. (One of the test publishers, CTB, uses five.) Like NAGB, all three develop achievement level descriptions before setting cut-scores. As is the case with NAEP, all three use multiple-choice and constructed-response questions. NAGB's process is distinguished by the presence of noneducation professionals on its panels, consistent with the legislative mandate to use a widely inclusive approach. Commercial efforts do not provide exemplary items for interpreting achievement level descriptions, as NAGB does.

States With NAEP-Like Performance Standards, **Jeffrey M. Nellhaus, Massachusetts Department of Education**

The Nellhaus study suggests that the States also have been influenced by the NAGB process. At least 23 States have developed achievement level categories similar to NAGB's. Most States have used one of three methods to set standards: modified Angoff, bookmark, or booklet classification, which are the same methods used by commercial test publishers, who are the contracted test developers for many States. Eight States have adopted five (instead of four) levels, appearing to have divided the Below Basic category into two parts. Nellhaus also notes that in a limited study of nine States, the results of State NAEP and State assessment programs tend to be most consistent at the two middle levels (Basic and Proficient). In general, States report a higher percentage of students at the Advanced level and a lower percentage at the bottom level.

Review of Alternative Models for Developing Achievement Levels

A Survey and Evaluation of Recently Developed Procedures for Setting Standards on Educational Tests, **Mark D. Reckase, Michigan State University**

Reckase's review of possible standard-setting methods reveals that many of the procedures suggested over the past decade have been used in limited research studies only or merely described as possible procedures. All would need extensive further development to connect the method to the policy and content frameworks, to develop methods for reporting results, and to withstand the type of public evaluations that have been applied to NAGB's process.

For NAGB to evaluate the potential usefulness of alternative standard-setting methodologies and their possible applicability for NAEP piloting, Reckase suggests the application of the following criteria:

(A) That judges can set the standard they intend.
(B) That the tasks that judges are asked to perform are moderate in their cognitive complexity.
(C) That the cut-scores have acceptable standard errors of estimate.
(D) That the process is replicable.

Exercising these criteria, Reckase suggests that the methodologies that have the most potential for NAEP are:

- Anchor-based.
- Bookmark.
- Generalized examinee-centered.
- Multi-stage aggregation method.

Advantages and disadvantages of each are discussed. Reckase also suggests that a combination of these methods may yield the best overall process.

Options for Modifying NAGB's Current Achievement Levels

***Looking at Achievement Levels,* W. James Popham, University of California at Los Angeles**

Since NAGB has examined the procedure for setting its achievement levels over the years, analyzed the ways in which the standards-based reporting is used and interpreted by various audiences, and compared these methods to others under development or in use by States and commercial test publishers, NAGB may wish to consider modifications that would make the achievement levels even more useful in conveying where improvement in student achievement may be taking place within the current levels. Popham's paper discusses the impetus for such a consideration and reiterates Secretary Riley's suggestions to explore a variety of ways to "convey to the American people that, yes, we have high standards . . . but yes, also we're measuring improvement or the lack thereof in a useful way" (p. 160).

Although Popham suggests that no one option is the perfect solution, he offers five modification options "that appear to be likely contenders for change" and discusses the strengths and short-comings of each: (1) add one or more achievement levels; (2) divide the current levels into distinguishable, within-level reporting categories; (3) make Below Basic a NAGB-sanctioned reporting category; (4) relabel the existing achievement levels, especially Proficient; and (5) lower the scale-score ranges associated with one or more achievement levels.

Finally, he offers his own recommendation for a possible solution, a "clarification-focused strategy that, because of a refinement in NAEP's below-goal reporting categories, will make it possible for NAEP reports to be a cause of celebration, not sorrow" (page 181).

SECTION 2

Reporting NAEP by Achievement Levels: An Analysis of Policy and External Reviews

William Brown Brownstar, Inc.

November 2000

Reporting NAEP by Achievement Levels: An Analysis of Policy and External Reviews

William Brown

Reporting Practices for NAEP

In the 1970's and 1980's, The National Assessment of Educational Progress (NAEP) primarily focused on (1) developing an appropriate survey test of national achievement, (2) devising means to validly administer the test so that it appropriately represented the Nation, and (3) reporting the results so that information could be conveyed to the public on the status of achievement nationally and the change in status over time. Little, if any, attention was given to creating qualitative performance levels in those early days.

The Alexander-James study group, which reviewed NAEP and proposed revisions, hinted at the need for "feasible achievement goals," but offered no real direction or mandate for the creation of performance levels. The NAEP panel, which commented on the report, went further and called for "descriptive classification" of achievement. The concept of achievement classifications, however, was controversial from the start. The legislation in 1988 that called for "appropriate achievement goals" for each grade and subject assessed by NAEP was grounded in compromise between the opposition of the House of Representatives and the Senate that allowed such action.

The reporting practices for NAEP in the 1970's and 1980's used conventional, normative statistics, such as national averages, percentiles, and standard errors, to evaluate change. The types of scores reported by NAEP drew less attention than such issues as the accuracy of sampling and the development of test items that were informative to curriculum specialists. Building an appropriate assessment program and reporting the results to the professional community was more important in the early years than making evaluative judgments about the quality of American education.

Discontent With NAEP Reporting

After the establishment of the National Assessment Governing Board (NAGB) as the policy director of NAEP, the reporting of NAEP results became a policy issue. NAGB questioned whether the results of NAEP were reported in a manner that was informative and useful. NAEP results in the past were reported conservatively, beginning with a focus on individual items and whether the performance of the Nation had changed significantly for the items of interest. Later, there was more interest in whether the performance of the Nation had improved or declined, so aggregated scores were examined.

The statistical model for NAEP reporting was based on normative techniques, with attention to significant change in performance over time. As with normative models, the norm group was the standard for measuring status or change. NAGB, however, was more interested in the qualitative aspect of national performance. To NAGB, the important consideration was whether

performance in the Nation was up to challenging expectations of what should be the performance expectations for the Nation's students.

Members of NAGB believed that normative models could mislead the public if the average score of the national group did not reflect sufficient quality. NAGB believed there were procedures to establish performance expectations using input from a variety of interested stakeholders, such as teachers, principals, curriculum specialists, and business leaders. The process had been satisfactory for establishing licensure cut-scores, so it was assumed that such a process could be expanded to establish multiple performance levels that would be useful in reporting NAEP results. If NAGB were successful in establishing multiple performance levels based on the core knowledge of what students should know to be proficient, reporting the proportions of students achieving at each level would have the qualitative aspects that were absent from normative results. In addition, these performance levels could be used as a means to set meaningful goals for improving the Nation's educational status. It was believed by NAGB that NAEP results reported on meaningful performance levels would be more understandable by the public and more useful to those who make instructional and policy decisions. The establishment of performance levels and reporting NAEP on this basis alone would be an important step in improving the form and use of the National Assessment results.

Vinovskis (1998), a professor at the University of Michigan, devoted an entire section to "Developing NAEP Performance Standards" in the report titled, *Overseeing the Nation's Report Card—The Creation and Evolution of the National Assessment Governing Board (NAGB).* This thoroughly researched section reported the history of NAGB's actions to create the NAEP achievement levels and the reactions, both positive and negative, to this innovation for NAEP. The presentation shows that significant members of the profession were opposed from the start to establishing achievement levels for NAEP. There was also disagreement about the adequacy of standard setting models and the technical criteria that were relevant to assess how well standards have been set. NAGB, however, persisted in implementing its policy to report NAEP achievement by the proportion of students scoring in one of the three achievement level categories—Basic, Proficient, and Advanced. Thus far, achievement levels have been established for Reading, Mathematics, Science, U.S. History, Geography, Civics, and Writing. NAGB continues to be convinced that its policy of reporting performance by achievement levels has been more useful to the public and to policymakers than if results were reported normatively for NAEP.

Cognizant that considerable controversy remains over the setting of achievement levels for NAEP, NAGB has commissioned research on standard setting as part of the standard-setting contract. NAGB also has sponsored sessions to discuss the technical issues and to solicit input from other fields that engage in standard setting. However, the policy issue of having valid performance levels for NAEP remains controversial.

Establishing the NAEP Achievement Levels

NAGB continues to use as its legislative authorization for achievement levels; the authorization language states that: "The National Assessment Governing Board . . . shall develop appropriate student performance levels for each age and grade in each subject area to be tested under the

National Assessment." The legislation continues by calling for achievement levels that are (1) devised through a national consensus process; (2) reasonable, valid, and informative to the public; and (3) updated as appropriate. To provide guidance to those contracted to implement the process, NAGB set forth the following six principles (later called guidelines):

1. The process shall establish three thresholds—Basic, Proficient, and Advanced—with the following definitions:

Basic:	This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.
Proficient:	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.
Advanced:	This level signifies superior performance beyond proficient.

   Note: These policy definitions were adopted for the 1994 assessments. The 1992 NAEP policy definitions for achievement levels were more descriptive and criterion related.

2. The process for developing achievement levels shall be conducted in phases and shall be widely inclusive of persons nominated nationally, thus insuring an opportunity for a national consensus of panelists to occur.

3. The existence of performance levels will be incorporated into all significant elements of NAEP (e.g., subject area consensus, exercise development, and selection and assessment methodology), and the achievement levels shall be used to report NAEP results so long as they are reasonable, valid, and informative to the public.

4. NAGB will exercise its policy judgment in setting the levels after reviewing information and input from a wide variety of sources and resource groups.

5. The achievement levels shall be used as the initial and primary means of reporting NAEP nationally and in the Trial State Assessment of NAEP.

6. The achievement level-setting process shall be managed in a technically sound, efficient, and cost-effective manner, and shall be completed in a timely fashion. NABG developed a document, "Guidelines for Setting Achievement Levels," that clearly described its rationale for each guideline as well as the practices and procedures that it expected to be implemented for each guideline. The degree to which the NAGB guidelines have been attained could be considered as the criterion for assessing the resulting achievement levels. This is particularly so because there is little agreement among the profession as to a single best process for

standard setting. As Dr. James Popham noted in his concluding summary at the Joint Conference on Standard Setting for Large-Scale Assessments,

> "We should not be too hard on ourselves or look for a level of precision and accuracy that is not attainable by normals . . . If we proceed in a reasonable, professional, and rational way, we can come up with standards that will be accepted. These standards can be defended against critics and lawsuits . . ."

NAGB's policy for reporting NAEP results by achievement levels continues to be preferred by the Board, but the Board is charged in H.R. 4328 to respond to the findings by the National Academy of Sciences report that the achievement levels established for NAEP are "fundamentally flawed." An analysis of the criticisms, which have occurred from the first evaluation, "Summative Evaluation of NAGB's Pilot Project To Set Achievement Levels On NAEP," to the present (Stufflebeam, Jaeger, and Scriven, 1991), is the final focus of this report. The analysis will describe the criticisms of the model, the process, and the end result of setting achievement levels. In addition, interpretative issues relating to the use of achievement levels will be analyzed.

A Review of Reports Relating to NAEP Achievement Levels

Report 1:
Summative Evaluation of NAGB's Pilot Project to Set Achievement Levels on NAEP

Authors: Daniel L. Stufflebeam, Richard M. Jaeger, and Michael Scriven
NAEP Date and Subject: 1990 Mathematics
Funding Source: National Assessment Governing Board

Findings:

1. NAGB is broadly representative of educational stakeholders but has "too little expertise from the psychometric and evaluation communities" to meet their policymaking and test design responsibilities. (Process)

2. The test item pool was inadequate to cover the range of achievement desired. (Model)

3. A changing NAEP item pool across years may confound with a meaningful analysis of how performance is changing. (Model)

4. The achievement level definitions were vague, ambiguous, confusing, and not explicit at the margins of the performance intervals. (Process)

5. The Angoff model was poorly implemented at the first achievement level-setting site (Vermont) and improved somewhat at the replication site (Washington, D.C.). (Process)

6. The final replication study improved and was characterized as "organized and smooth." (Process)

7. The model called for analysis of data collected at mid-year but asked for performance estimates at the end of the year. (Model)

8. Panelists were instructed to assume that no guessing occurred. (Process)

9. The authors cited a study by Linn (Linn et al., 1991) that reported excessive between-panel variation between the Vermont and Washington panels. (Process)

10. Extensive work is necessary to improve the NAEP item pool and the methodology for the achievement level setting. (Process)

11. Consideration should be given to alternative methodological strategies for securing judgments of appropriate achievement levels. (Model)

Report 2:
The Validity and Credibility of the Achievement Levels for the 1990 National Assessment of Educational Progress in Mathematics

Authors: Robert L. Linn, Daniel M. Koretz, Eva L. Baker, and Leigh Burstein
NAEP Date and Subject: 1990 Mathematics
Funding Source: National Center for Education Statistics

Findings:

1. In the Vermont achievement level panel, some items that were common across grades lacked a coherent progression of difficulty across grade levels for the three achievement levels. (Process)

2. A systematically lower bias in performance levels was introduced when some panelists who participated in Vermont decided not to participate in the second replication in Washington, D.C. (Product)

3. The subgroups of panelists within a grade had significant variation in their ratings for Basic. (The variations for Proficient and Advanced were not presented.) Therefore, the achievement levels will be subject to variation if set by different groups. (Process)

4. The correlation is high between the performance level ratings for items and the *p*-values of these items. Therefore, the achievement levels are substantially affected by normative considerations. (Model)

5. The expectation for performance as estimated by panelists was not commensurate with the differences in difficulty of the content areas within the mathematics area. (Product)

6. The exclusion of higher order and estimation items from the final calculation of the achievement levels was not made known to the panelists and is a deviation from an acceptable practice in standard setting. (Process)

7. Too few students are classified as Proficient or Advanced without corroborative evidence from independent sources. (Product)

Report 3:
Interpretations of NAEP Anchor Points and Achievement Levels by the Media in 1991

Authors: Daniel Koretz and Edward Deibert
NAEP Date and Subject: 1990 Mathematics
Funding Source: National Center for Education Statistics

Findings:

1. The presentation of NAEP results in mathematics in 1991 by the media was simplified to the point of being misleading to the public regardless of the provision of NAEP anchor points or NAEP achievement levels. Both processes were more informative than NAEP scale scores, which have only an average score for a point of reference. (Interpretation)

2. The NAEP results reported by anchor points were misleading because of vague language about "grade level" as were the achievement level results on what students should know at each achievement level.

3. The press generally ignored or vastly simplified the results presented as *p*-values for items regardless of reporting format (i.e., anchor levels or achievement levels). (Interpretation)

4. The results of anchor levels and achievement levels were interpreted by the media as discontinuous performance rather than points on a continuous scale. (Interpretation)

5. Reporting NAEP results is a complex process and is unlikely to be interpreted properly by press writers without specific guidance in the appropriate wording of releases. (Interpretation)

Report 4:
Survey of Reactions to the Use of Achievement Levels in Reporting 1990 NAEP Mathematics Results

Author: Aspen Systems Corporation
NAEP Date and Subject: 1990 Mathematics
Funding Source: National Assessment Governing Board

Findings:

1. National education and policy groups:
   a. Found the NAEP reporting by achievement levels to be very useful (46.4%). (Interpretation)
   b. Found the achievement levels to be clear in conveying significance of student performance (50% had clarity of achievement levels rated as 7 or higher on a 10-point scale). (Interpretation)
   c. Preferred three levels of reporting (Basic, Proficient, and Advanced) to two levels of competence. (Interpretation)

2. State education advisors to the Governors:
   a. Considered the NAEP achievement levels to be very useful (61%). (Interpretation)
   b. Found the achievement levels to be clear in conveying the significance of student performance (72.5% had clarity of achievement levels rated as 7 or higher on a 10-point scale). (Interpretation)
   c. Preferred three levels of reporting (Basic, Proficient, and Advanced) to two levels of competence (75.6%). (Interpretation)

Report 5:
Setting Performance Standards for Student Achievement: A Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels

Authors: The National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels—Lorrie Shephard, Principal Investigator; Robert Glaser and Robert Linn, Panel Chairmen
NAEP Date and Subject: 1992 Mathematics and Reading
Funding Source: U.S. Department of Education

Findings:

1. The Angoff method requires panelists to make conceptual judgments that are complex: (1) they must conceptualize borderline performance specifically enough to make *p*-value estimates; (2) they must be able to understand the relationships among the content framework, the achievement level descriptions, the NAEP assessment items, and the performance levels; and (3) there should be evidence that panelists are grounding their estimates to factors in the achievement level descriptors and to the NAEP items. The study concluded that the achievement level descriptions evolved significantly during the process of setting achievement levels. The panelists reported that personal and experiential background influenced their judgments as well as the achievement level descriptions. As a result, the study concluded that "many participants were not making systematic judgments based in specific features of the descriptions." (Model)

2. The study established as a criterion of validity an expectation that judges should be internally consistent in making item judgments except where reasonable reasons exist for varying their ratings, such as estimating higher ratings for items that should be taught but are not taught now. The study concluded that this expectation for consistency of ratings was not met in the following instances: (1) when comparing cut-scores recommended from dichotomous or polytomous items, and (2) when cut-score estimates were not consistent when comparing items that were relatively easy or hard. (Model)

3. The study also looked for inconsistency in judges' ratings of different cognitive processes, such as numerical operations, geometry, and algebra. The study concluded that no systematic variations occurred because of the dimensions of the assessment. (Model)

4. The study found that the judges became more consistent across rounds of ratings. (Model)

5. The study found that judges could estimate three cut-scores on an assessment, either concurrently or serially, without a significant magnitude of difference. (Model)

6. The study postulated that allowing judges to evaluate intact test booklets would provide "more complete and integrated evidence of student performance," which would lead to differences with cut-scores set by the Angoff method. The results of the study were

inconsistent. Whole-booklet reviews led to higher cut-scores for the Basic level and lower cut-scores for the Advanced level. (Model)

7. The study concluded that arriving at a consensus among the judges was a matter of averaging the scores of judges, who demonstrated great variation across each round of rating rather than converging to a shared cut-score. (Model)

8. The study concluded that an analysis of differences among groups, teachers, nonteachers, and the public, did not reveal systematic group differences. (Model)

9. The study was decidedly critical of the Angoff method based on the author's assertions that:
   a. The use of Angoff method becomes "murky" when items span a range of content in item difficulty and the level estimated is more than a "must know to be minimally competent" category. (Model)
   b. An item-by-item approach does not allow consideration of particular combinations of performance in arriving at achievement levels, as is the case with a whole-booklet approach. (Model)

Report 6:
The Trial State Assessment: Prospects and Realities

Authors: The National Academy of Education Panel—Robert Glaser and Robert Linn, Panel Chairmen
NAEP Dates and Subjects: 1992 Mathematics and Reading—Grade 4
1992 Mathematics—Grade 8
Funding Source: U.S. Department of Education

Findings:

1. Citing a previous NAEP study, "Setting Performance Standards for Student Achievement," the authors of this report conclude:
   a. "Judges . . . were unable to make consistent judgments when translating their substantive expectations into cut-scores on the NAEP scale." (Model)
   b. Inconsistencies occurred more in judges' estimations on multiple-choice/short-answer questions than on substantive dimensions of the assessment (constructed-response items). (Model)
   c. Item-by-item difficulty judgments (Angoff method) are not adequate procedures for arriving at a cut-score. (Model)
   d. Three alternative approaches should be explored for setting NAEP achievement levels: (1) contrasting-group, field-based studies, (2) an item-mapping procedure, and (3) a total student performance (whole-booklet evaluation) procedure. (Model)

Report 7:
Quality and Utility: The 1994 Trial State Assessment in Reading

Authors: The National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: 1994 Trial State Assessment in Reading, Robert Glaser and Robert Linn, Chairmen
NAEP Date and Subjects: 1992 Reading and Mathematics
Funding Source: U.S. Department of Education

Findings:

1. The study repeated the panel's concerns expressed in a previous study. These concerns were: (1) the levels are not internally consistent or coherent, (2) the proportions of students scoring above the levels do not appear to be reasonable, and (3) any item-by-item procedure for setting achievement levels is inadequate. (Model)

2. The study reviewed NAGB's research effort to determine if panelists could review the NAEP items and correctly match them to an achievement level category. The analysis of the research results indicated that the expected patterns emerged: Higher proportions of students correctly answered the Basic items than items in the other categories and Proficient items were answered correctly more often than Advanced items. The pattern was the same at each grade level. However, the panel, remained concerned about the variation in individual items that judges had classified in each category. (Product)

3. The study also reviewed NAGB's research on how well students performed at the three levels and whether the student performance agreed with the skills included in the achievement level descriptions. (Process)

Report 8:
Validating Inferences From National Assessment of Educational Progress Achievement-Level Reporting

Author: Robert L. Linn
NAEP Date and Subjects: 1994 Geography, U.S. History, and Reading
Funding Source: National Center for Education Statistics

Findings:

1. Citing previous research by Shepard et al. (1993), the author contends that item format (right-wrong answer or partial-credit constructed response) significantly confounds the setting of achievement levels. The cut-score depends largely on the proportion of items of each type and NAEP tests vary in item format proportions. (Model)

2. The author contended that the initial set of achievement level definitions in force in 1991 was extensive, and these definitions described specific outcomes or performances expected of students. These achievement level performances have not been validated empirically. A later set of policy definitions adopted in 1994 were streamlined, and these definitions have considerably lessened the issue of validly interpreting the policy definitions. (Product)

3. The author cites problems in using the NAEP achievement levels to make substantive interpretations of what students in a particular level can actually do. He concludes that the ability to validate the student performance in achievement levels is lacking. (Product)

Report 9:
Grading the Nation's Report Card

Authors: James W. Pellegrino, Lee R. Jones, and Karen J. Mitchell, Editors
NAEP Dates and Subjects: 1990, 1992, 1994, and 1996 Reading, Mathematics, Geography, U.S. History, and Science
Funding Source: U.S. Department of Education

Findings:

1. The study revisited the collective list of criticism of previous reports criticizing NAGB achievement levels.

2. The study extensively described the process used to set achievement levels on the 1996 Science assessment and concluded that the process used in setting the achievement levels resulted in levels that NAGB could not accept as reasonable. After considering the recommendations from the achievement level setting panelists, NAGB authorized additional study and subsequently modified seven of the nine cut-scores. When the science results were released, they were reported as "What Do Students Know?" (Process/Model)

3. The study cited a number of perceived benefits made possible when reporting by achievement levels. Even so, the authors believe there are serious failings in the current process, and they encourage NAGB to continue to search for more valid and useful ways to report achievement by standards. (Model)

4. The aberration in process used in setting the science achievement levels was not well described, and the authors speculated about future decisions that may be made by NAGB and what factors may affect these decisions. (Model)

5. The study stated that the achievement level-setting model is flawed because: (1) too few students are classified as Advanced for the level to be believable, (2) the item format and difficulty of the items have a confounding effect on the levels, and (3) the panelists have difficulty in estimating the *p*-value correct for an item and are likely to interject a systematic bias that underestimates high probabilities and overestimates low probabilities. (Model)

6. The authors believe that a tighter alignment of the NAEP items and the preliminary achievement level descriptions is important and necessary. To accomplish this, they believe that the preliminary achievement level descriptions should be available and made an integral part of the development of assessment items. (Process)

7. The authors affirmed that NAGB has the authority and responsibility to adopt achievement level cut-scores. However, technical expertise is available to NAGB and the technical experts should have a role in establishing and reviewing the process and the results of the achievement level-setting process. (Process)

8. The authors believe that consequences data and, when possible, comparative data from external sources should be provided early in the achievement level-setting process. (Process)

9. The authors acknowledge that no method currently exists that is clearly proven as a better approach than the item-by-item (Angoff) method. However, they propose research on a model that includes: (1) judgments based on aggregate student performance, (2) examination of comparable referent data, and (3) a process that involves policymakers and educators in the final standard-setting process.

Report 10:
The Validity of the 1992 NAEP Achievement Level Descriptions as Characterizations of Mathematics Performance

Authors: Leigh Burstein et. al.
NAEP Date and Subject: 1992 Mathematics
Funding Source: National Center for Education Statistics

The summary finding of the study was that the published content descriptions do not validly characterize what students within specified levels can do. This conclusion was based on analyses indicating that: (1) the performance on exemplar items varies from reasonable expectations for some items, (2) the definitions of achievement levels overlap considerably and differ in minor or unclear ways, (3) the 1992 NAEP Mathematics items provided inadequate coverage at some achievement levels, and (4) the performance of students or items categorized at various achievement levels was lower than could be reasonably expected in some instances. (Product)

Categories of Criticisms and Responses

Model Criticisms

1990 Mathematics

- The NAEP test items were inadequate to cover the desired range of achievement. (Report 1)
- A changing NAEP item pool may confound meaningful trend analysis. (Report 1)
- The model uses midyear data but asks for end-of-year performance estimates. (Model)
- Consideration should be given to alternative methodological strategies for securing judgments. (Report 1)

Reaction:

- NAEP made improvements in the 1992 item pool that improved coverage and NAEP item development continued to improve in other subjects.
- The changing item pool is addressed by the item response theory model used for NAEP analyses.
- NAGB has authorized the testing of a variety of methodological refinements for the modified Angoff method and has adopted some of the better refinements in its current process.

1992 Mathematics and Reading

- The Angoff judgments are too complex (Report 5) because
  (1) must conceptualize borderline performance for three levels
  (2) panelists must understand the relationships among the content framework, the achievement level descriptions, and performance on the NAEP items; and
  (3) judgments by panelists must be grounded in factors directly related to the achievement level descriptions.
- Panelists should be internally consistent in making judgments across items of various formats and types; there is evidence that panelists are not able to do this. (Reports 5, 6, and 7)
- Allowing panelists to evaluate intact test booklets was purported to be better than the item-by-item judgments. (Report 5)
- Achieving consensus across panelists was not evident; the consensus was only an average of discrepant panelists. (Report 5)
- The use of Angoff method becomes “murky” when the estimate is for three levels and not just a minimum score for a group. (Report 5)
- Item-by-item ratings do not allow for consideration of various combinations of performance. (Reports 5, 6, and 7)

- Alternative models should be considered, such as contrasting groups, item mapping, and whole booklet. (Report 6)

- Student performance in the achievement levels did not match NAGB descriptions very well. (Report 7)

- The model was flawed because: (1) too few students were classified as Advanced to be believable, (2) item format was a confounding factor, and (3) panelists have a systematic bias when estimating item by item. (Report 9)

- Research on new and different models should be conducted. (Report 9)

Reaction:

- ACT refined its training process substantially in 1994 to ensure that panelists were able to make distinctions at the borderline of the achievement level descriptions. In the 1996 science achievement level-setting process, panelists wrote borderline descriptions as part of the process.

- ACT training for panelists is integrated into the 5-day training and rating sessions used in standard setting. Panelists indicate on their evaluations that they have confidence in their understanding of the process.

- One criticism of the Angoff model contended that item-by-item ratings do not allow for compensatory performances on various combinations of items. Although it is accurate that item-by-item judgments are made independently, the Angoff model has compensatory aspects because the total score required can be obtained from any combination of items.

- ACT is aware that panelists make decisions that cause cut-scores to be different for various item formats. ACT has studied this phenomenon extensively but does not have a satisfactory answer as to why it occurs. ACT uses Reckase charts to provide as feedback between rounds. This feedback seems to lessen the format effect but does not eliminate it. Panelists are aware of the effect and continue to produce estimates that yield different cut-scores for different item formats. Panelists' ratings continue to be different even when they are aware of the differences.

- There are repeated criticisms of the low percentage of Advanced students that result from the NAEP achievement levels. These low percentages are not perceived as either reasonable or believable. However, panelists were provided consequences data during the process, and these data had little effect on the panelists' judgments of how high to set the standards for what students should know to be congruent with the achievement level descriptions.

- ACT has researched the whole-booklet procedure and found that this method consistently produces higher, not lower, cut-scores. If the Angoff method is faulted for achievement levels that are too high, the use of a whole-booklet procedure also would be faulted.

- The ratings of panelists converged after round one, and there were no substantial differences after the final round in the average cut-scores of teacher, nonteacher, and public panelists.

- ACT has explored the contrasting-groups procedure in research studies. This procedure has the logistical problem of having to be conducted at nearly the same time as the testing, and it is expensive to conduct. ACT studies indicate that teachers estimate their students to be more competent in relation to achievement level descriptions than the students' performance ultimately reflects.

- ACT conducted several research studies related to the use of item maps. The item-mapping procedure was rejected because of problems associated with choosing a response-probability criterion value. In several trials, cut-scores from item mapping were similar to those produced by the Angoff method.

- The Reckase charts were adopted as feedback to provide panelists with information that relates the performance on the item to cut-score projections.

Process Criticisms

1990 Mathematics

- The achievement level definitions were vague, confusing, and not explicit at the margins of the achievement intervals. (Report 1)

- The Angoff model was implemented poorly. (Report 1)

- Panelists were instructed to ignore guessing. (Report 1)

- There was evidence of excessive variation between panels. (Report 1)

- Extensive work is necessary to improve methodology. (Report 1)

- Common items across grades lacked coherent progression in performance. (Report 2)

- Panelists of various types within a grade had significant variation. (Report 2)

- Deletion of some items from final calculations was a problem. (Report 2)

- Criticisms from previous reports were revisited. (Report 9)

Reaction:

- The policy definitions for Basic, Proficient, and Advanced were changed substantially before the 1994 achievement levels were set. Current policy definitions focus descriptively on the Proficient level and are less criterion referenced.

- The initial problems in the process for setting achievement levels were recognized and a contract for standard setting was awarded to ACT, a recognized leader in standard setting for licensure exams.

- Although there is no explicit correction for the effect of guessing, its influence is one factor that is discussed in the training of panelists. Panelists are instructed about chance probabilities, and the Reckase charts reveal guessing effects.

- To check on consistency among panelists, ACT divides the grade panels into two rating groups as a routine practice. Differences in the cut-score recommendations of the two groups are small. There are no significant differences in the ratings for items that are common to both groups and there is no significant difference in the ratings or by panelist type.

1992 Mathematics and Reading

- The collective list of process criticisms from 1990 was also referenced to the 1992 achievement levels. (Report 9)

1994 Geography, U.S. History, and Reading

- There should be a close alignment between the NAEP achievement level descriptions and the NAEP item development process. (Report 9)

- NAGB should be encouraged to involve additional expertise in the final standard-setting process. (Report 9)
- NAGB should be encouraged to direct the use of consequences data and external references early in the achievement level-setting process. (Report 9)

Reaction:

- Efforts have been made during training sessions to align the panelists' perceptions of the content framework and the achievement level descriptions. The item pool also includes a broader range of item difficulties than previously. A close alignment of NAEP items and achievement level descriptions remains a problem because achievement level setting comes after item development.

- ACT has proposed, and NAGB has accepted, advice to explore the effect of providing consequences data in various forms and at various times during the process. To date, studies show that consequences data have limited impact on panelists' perception of what students should know.

- Providing valid and comparable external performance referents is more problematic, especially if it is to be done across all subjects. Some external referents, such as the percentage of advanced placement (AP) students for a given subject who score at or above the level required to earn college credit, seem to be applicable as input into the achievement level-setting process. But, on reflection, it is obvious that the motivation to do well on these tests is far different than the motivation for a low-stakes test such as NAEP. There also is the issue of direct versus indirect instruction. Students are instructed for a year specifically for

the AP exams, but the NAEP frameworks are not from a universal curriculum, therefore the extent to which students have been instructed on the content in the NAEP framework is unknown. If a decision were to be made to use external referents, could the use of referents be standardized across subjects? If not, a serious inconsistency would be introduced in the achievement level- setting process.

- NAGB has access to considerable technical expertise through members of the Board, the technical staff of ACT, the Technical Advisory Committee on Standard Setting advisory panel, and representatives of National Center for Education Statistics and Educational Testing Service. After application of the technical advice available, NAGB is the final decisionmaker in setting NAEP standards.

Product Criticisms

1990 Mathematics

- A systematic lower bias was introduced in the process when some Vermont panelists did not participate in the following session in Washington. (Report 2)

- Panelists' ratings correlated highly with normative data points, so the results must be affected by normative considerations. (Report 2)

- Too few students were classified as Proficient or Advanced without corroborative evidence. (Report 2)

Reaction:

- The lower bias from defecting panelists was a one-time aberration in the process used in 1990.

- It is assumed there will be a positive relationship between performance estimates from normative data and performance estimates from a standard-setting process. An absence of such a relationship would indicate that something was invalidating the rating process. The existence of a relationship does not mean that panelists were being influenced by normative considerations.

- Corroborative evidence for validating the achievement levels is tenuous because the levels are based on: (1) student performance where the motivation is questionable, (2) student performance on high-stakes tests such as AP tests that are directly influenced by instruction, and (3) comparisons to international tests that have completely different content frameworks and sampling plans.

1992 Mathematics and Reading

- The performance of students was lower on the achievement levels than would be reasonable. (Report 10)

Reaction:

- The NAEP achievement levels have been criticized for identifying too few students at the Advanced level. This may be the prime criticism of the process but is actually more a policy issue than a technical problem. If there are no valid external referents for comparison and if studies have shown that teachers generally overestimate the performance of their upper level students, what can be used to determine the reasonableness of a standard? Policymakers could, and sometimes do, use political viability as a factor in adopting the final cut-score. Another option would be the use of a normative model to define the percentage of students who should be advanced. But these solutions would not relate the achievement levels to the achievement level descriptions very well. If the panelists who set the levels have the consequences data early enough in the process to make substantive changes, the resulting recommendations from these panelists should be reasonable within the context of standard setting. If other outcomes are desired, the principles undergirding the process should be reconsidered.

1994 Geography, U.S. History, and Reading

- The validation of the achievement level descriptions has not been validated.

Reaction:

- Procedures for achievement level setting have been improved and stabilized across subjects. Levels appear to be based on what students know rather than what they should know.

- Predictive validity for what students should know at each achievement level may not be possible to demonstrate in traditional psychometric terms.

1996 Science

- The process used to set science achievement levels produced levels that were very high standards. NAGB modified seven of the nine cut-scores. The resulting levels appear to be based on what students know rather than what they should know.

Reaction:

- The achievement level-setting process for science was similar to those used in other subjects, yet the results were different. The final stage for setting achievement levels resides with NAGB. In this one case, NAGB exercised its discretion to change the levels to be more in line with other subjects and to be judged as reasonable to persons using the levels in policymaking.

Actions Taken by NAGB to Bolster the Achievement Level-Setting Process

Since the 1992 achievement level-setting process was established for reading and mathematics, NAGB has engaged, through its contract with ACT, in extensive review and research of its achievement level- setting process. Advice was sought from technical resources that addressed concerns raised internally as well as those raised by external evaluations or groups. ACT maintains a Technical Advisory Committee on Standard Setting as well as an internal technical

advisory team. The purpose of these groups is to advise NAGB and ACT on technical issues or study design. The refinements made by NAGB after 1992 included the following actions.

1. Articulation of NAEP Curriculum Frameworks, Item Development, and Achievement Level Descriptions
   NAEP items are developed using the NAEP content framework as a guide for content. The NAGB achievement levels also interface with item development more than in 1990 or 1992. Noticeable improvements have been made in having an adequate item distribution and having sufficient items that can serve as exemplars for each of the achievement levels. Because the achievement levels are not known at the time of item writing, estimations must be made during item development regarding item difficulties needed. Additional emphasis should be given to correlating the NAEP content framework, the NAGB policy definitions, and the achievement level descriptions. Even so, there has been improvement in the adequacy of items for each level.

2. The Sampling Plan
   NAGB's policy on panel representativeness has steadfastly insisted on a process that resulted in panels composed primarily of teachers who taught appropriate grades and subjects as well as knowledgeable educators who are not teachers. Representatives of the public must have some expertise in the content tested and experience with the education of students from appropriate grades. The process includes a group of nominators who represent the diversity of the Nation. These nominators make recommendations for panelists. The panel selected is proportionally representative of the categories of persons sought, the regions of the Nation, and appropriate ethnic and gender proportions.

   Questions were raised by several evaluations about interrater consistency, especially with such different backgrounds. Could they converge to a point of consensus in their ratings? When panelists from the standard-setting sessions were compared, the type of panelist (e.g., teacher, nonteacher, public), was not a factor in whether achievement levels were set high or low. In most cases, differences among these groups were minor and did not present a consistent pattern.

3. Extensive Training of Panelists
   A primary criticism of the use of the modified Angoff method was the complexity that panelists faced when applying the model. Assertions were made that the conceptual complexity of the model confused panelists and rendered their judgments invalid. After ACT was awarded the contract for standard setting, they made the training for panelists more extensive. Currently, 5 extensive days are used for training panelists and the setting of achievement levels. The training begins with developing a common understanding of the purposes of standard setting and the general procedures that will be followed. The process continues with a familiarization of the panel with the NAEP test that was administered and the protocols used in scoring. The process has differed somewhat by subject area, but an orientation includes the NAEP content frameworks and the achievement level descriptions. Until recently, the process allowed some modifications in the achievement level descriptions if needed to add clarity.

The training also focused on the development of a concept of borderline performance for each of the three performance levels. Writing descriptions of borderline performance is now part of the achievement level-setting process. Training also was provided on how to estimate performance for dichotomous and polytomous item formats. After the contractor has confidence that the panelists are well trained and clear about the rating tasks, the item-rating process begins. Each standard-setting session, including the training component, is piloted and revisions are made before the actual standard-setting session begins. When surveyed panelists consistently report their satisfaction with the training and the time allowed for the rating process.

4. Pilot Studies and the Research Agenda

ACT has consistently examined the efficacy of a variety of modifications that have been suggested by technical resource groups. A partial list of modifications tested includes:

a. A variety of approaches to secure more nominations and better participation rates for panelists.
b. A comparison of achievement levels set by item-by-item ratings versus panelists making a composite rating for each block of items.
c. A comparison of ways to rate constructed-response items. One comparison required panelists to review student papers and select three papers that represented borderline performance for each achievement level. Other procedures required panelists to estimate the mean score of each borderline group on the constructed-response items or the proportion of each borderline group that would score at each score point. A hybrid model that was also studied required panelists to select papers to represent the cut-scores in round one and use the mean estimation procedure for subsequent rounds.
d. Studies to determine the usefulness of feedback to panelists that provides them with "whole booklets" of students' performance that were selected to be at or near the borderline performance for the achievement levels.
e. Research on the effect of consequences feedback on panelists' ratings. Variations included:
   (1) Feedback on the percentage of students who would score at or above a cut-score based on the cut-score established from round two ratings.
   (2) Feedback on the percentage scoring at or above achievement levels set during the final round.
   (3) Comparisons of the effect of timing the consequences data early or late in the rounds of achievement level setting.
f. Studies of the impact on panelists of the provision of interrater consistency graphs as feedback.
g. Studies testing a process called the Item Difficulty Categorization Procedure, which required panelists to determine the consistency between what students within an achievement level can or cannot do as compared to what they should be able to do.
h. Studies to determine differences between the ratings of panelists who rated the three achievement levels with two points of reference, what students

should know and a second rating on how they would score on a NAEP test. This was to determine if panelists distinguish between estimated performance on an actual test and an estimate of what students should know.

i. Studies evaluating separate item-rating methods for items that are scored right/wrong and constructed-response items with partial credit scoring. The preferred method for estimating multiple-choice items was to focus on borderline student performance, estimating the percentage of students at the lower borderline of achievement who would answer the item correctly. The preferred method for constructed-response items was to have panelists estimate the average score on each item for students performing at the borderline of the achievement level.
j. Studies of the most informative method of providing feedback to panelists on their consistency of item ratings with other panelists.
k. Testing the effect of various forms of feedback between rounds of ratings. For example, consequences data showing actual student performance were incorporated into the process in various ways and as early or late feedback. Information was also provided to panelists that allowed them to review their estimated performance for each item in relation to the cut-score associated with the estimated item performance (Reckase charts). This allowed panelists to know the effect of their performance ratings by item or item type.
l. Various alternative models investigated by ACT for standard setting, including: (1) item mapping, (2) item score string estimation, (3) whole-booklet method, (4) Reckase method (as a standard-setting model), and (5) the grid model (for writing). The models tested thus far have not produced convincing evidence of their superiority to the modified Angoff method used in setting NAEP achievement levels.

At the same time as ACT implemented the Angoff method of standard setting, it evaluated other standard-setting methods. These studies indicated that other standard-setting methods have inadequacies and/or differences that make them problematic as viable replacements to the model currently used by ACT to set NAEP standards. Through the evaluation of other standard-setting models and the testing of alternative procedures that could be used with an item-by-item rating process, ACT has made considerable modifications to the process usually associated with the Angoff method. These refinements provide considerable improvement over the processes used at the start of the NAEP achievement level-setting process.

Summary

Nearly a decade ago, NAGB determined that reporting the NAEP results would be more meaningful and useful if NAEP achievement levels were established. Reporting results by the proportion of students who attained Basic, Proficient, or Advanced levels would replace normative reporting such as national averages, quartiles, or item *p*-values. NAGB's decision was embroiled in controversy then, and it remains so today. Most groups acknowledge that there is

popular support for achievement levels, even challenging ones, as the NAEP levels have proven to be.

The Policy Issue

The policy issue for NAGB continues to be whether NAEP results should be interpreted in terms of quality through achievement levels or normatively in terms of derived norms representing achievement status. With the question of educational quality in the nation very problematic at the beginning of the 90's, NAGB was clearly troubled by reporting on the average score of the Nation as the referent of quality. NAGB believed that qualitative reporting would provide an impetus for change even if the performance levels of students were not satisfactory initially. The policy decision to establish performance levels proved to be as technically complex as it was forward thinking. During the past decade, NAGB has persisted with its policy decision, even in the face of considerable criticisms from noted psychometricians who labeled the achievement level-setting process as flawed. In response to the criticism from reviewers and in search of improvements to the achievement level-setting process, NAGB continued to study alternate procedures that might improve the modified Angoff method that was in use. A complete description of these studies, conducted by ACT as part of the standard-setting contract, was beyond the purpose and scope of this section. However, each study is available through NAGB and/or ACT.

The results of the research conducted by ACT have improved the standard-setting model considerably. Changes in the panelists' training have been notable. The additional sources of input for the various rounds of ratings and the use of participant feedback on the process have made today's standard-setting process much more refined and satisfactory than was the case at the beginning of the process. Much of the criticism in the literature, however, is of the process used in the first and second standard-setting sessions. Clearly the problems that were identified with the initial process were concerns to be addressed, and many of them have been studied. The question to be addressed now is whether the recommendations by critics to abandon the current achievement level-setting process is warranted by the problems identified. The next question is whether there is a viable and tested model available that will produce more valid, more reasonable results.

The successful implementation of an untested standard-setting process to set NAEP achievement levels would be highly problematic. Unless a process can be demonstrated to be more appropriate than the item-by-item rating process used by ACT to set achievement levels, change should not be considered.

References

Aspen Systems Corporation. (1992) *Survey of Reactions to the Use of Achievement Levels in Reporting 1990 NAEP Mathematics Results,* prepared under contract with the National Assessment Governing Board.

Burstein, L. et al. (1993) *Describing performance standards: Validity of the 1992 National Assessment of Educational Progress achievement level descriptors as characterizations of mathematics performance*. Educational Assessment 3:9–51.

Glaser, R. and Linn, R. (1993) *The Trial State Assessment: Prospects and Realities*. Washington, DC: The National Academy of Education.

Glaser, R. and Linn, R. (1996) *Quality and Utility: The 1994 Trial State Assessment in Reading*. Washington, DC: The National Academy of Education.

Koretz, D. and Deibert, E. (1993) *Interpretations of NAEP Anchor Points and Achievement Levels by the Media in 1991*. Pre-edited Draft, Institute on Education and Training. Washington, DC: RAND.

Linn, R.L. (1998) *Validating inferences from national assessment of educational progress achievement-level reporting.* Applied Measurement in Education 11:23–47.

Linn, R.L., Koretz, D.M., Baker, E.L., and Burstein, L. (1991) *The validity and credibility of the achievement levels for the 1990 National Assessment of Educational Progress in mathematics.* Los Angeles: University of California at Los Angeles.

Pellegrino, J.W., Jones, L.R., Mitchell, K.J. (1999) *Grading the Nation's Report Card*. Washington, DC: National Academy Press.

Shepard, L., Glaser, R., Linn, R. (1993) *Setting Performance standards for student achievement: A Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels*. Washington, DC: National Academy Press.

Stufflebeam, D., Jaeger, R., and Scriven, M. (1991) *Summative Evaluation of NAGB's Pilot Project to Set Achievement Levels in NAEP*. Ann Arbor, MI: College of Education, University of Michigan.

Vinovskis, M.A. (1998) *Overseeing the Nation's Report Card: The Creation and Evolution of the National Assessment Governing Board (NAGB).* Ann Arbor, MI: Institute for Social Research, School of Public Policy, University of Michigan.

SECTION 3

A Survey and Evaluation of Recently Developed Procedures for Setting Standards on Educational Tests

Mark D. Reckase Michigan State University

November 2000

A Survey and Evaluation of Recently Developed Procedures for Setting Standards on Educational Tests[1]

Mark D. Reckase

According to *Webster's New Collegiate Dictionary* (1977), a "standard" is "something set up and established by authority as a rule for the measure of quantity, weight, extent, value, or quality." In the case of achievement levels for the National Assessment of Educational Progress (NAEP), the authority is the National Assessment Governing Board (NAGB), and the standard is for assessing the quality of students' achievement in specified areas. Although the dictionary definition of a standard gives a general sense of what a standard is, a definition that is more specific to the testing situation may provide a useful supplement. Cohen, Kane, and Crooks (1999) suggest the following definition as a guide to their new procedure for standard setting. A standard is an "explicit decision rule that assigns each examinee to one of several categories of performance based on his or her test score." The two definitions together provide a general framework for discussing standard setting in the context of NAEP.

Although the definitions tell what a standard is, they do not provide any guidance as to how the "authority" should go about establishing the standard. The purposes of this section are to provide a framework for considering alternative methods for establishing standards and then to review some of the newer methods that have been suggested for setting standards. These methods will be considered in light of the constraints that are imposed by the data collection requirements for the NAEP.

Structural Components of a Standard-Setting Process

Standard-setting methodologies usually contain the following five components: (1) an "authority" to set policy, (2) a content domain that is the focus of the standard, (3) a selection of persons to make judgments about desired levels of performance, (4) methodology for collecting judgments and estimating standards, and (5) some means for reporting the results of the process. Each of these components will be elaborated upon to set the stage for discussions of the standard-setting process.

Authority and Policy

The first component in a general framework for standard setting is suggested by the fact that standards are set by "authority." This means that there is an agency that calls for the existence of the standard and provides a policy definition for the standard. For NAEP, the agency is NAGB and the policy is given in NAGB (1990).

The policy provides definitions for three achievement levels (Basic, Proficient, and Advanced) and general guidelines for the process used to estimate cut-scores between the achievement level

[1] This paper was written on contract to the National Assessment Governing Board. All opinions expressed in the paper are solely those of the author.

categories.[2] The guidelines include the type of standard-setting procedure to be used, the types of individuals that should be involved in the standard-setting process, and other information about how the standards will be used in reporting NAEP results. The "authority" of an agency guides all standard-setting procedures and its policy. Documentation for a standard-setting method should be clear about the agency and any policy issues that guide development of explicit decision rules.

Content

The second structural component of a standard-setting process is the content domain on which the standard is to be set. That is, standards must be set on something, and the domain tells what the "something" is. In the context of NAEP, the domain is the content that is assessed by a specific NAEP assessment (e.g., mathematics, writing, and geography). In some cases, the domain is defined by the content of the test that is the target for the standard-setting operation. More often, however, the domain is being described in detail in a framework or content standards document (e.g., NAGB, 1998). Furthermore, the policy definitions for standards are often translated into content-specific descriptions to guide the formal standard-setting process. The ACT/NAGB process in current use for NAEP now contains an explicit step for translating policy definitions into content descriptions (ACT, 1997). Other standard-setting processes also create formal content descriptions to guide the process (e.g., Cohen, Kane, and Crooks, 1999; Kahl et al., 1995). Some processes allow the test and those involved in the standard-setting process to accept the domain implied by the test content without a formal description (e.g., Sireci, Robin, and Patelis, 1997).

Judges

The third structural component in a standard-setting process is the selection of the judges who will translate the policy definitions and content domain descriptions into decision rules that are the end result of the process. Virtually all people working in the area of standard setting acknowledge that standard setting is a judgmental process (Jaeger, 1989; Pellegrino, Jones and Mitchell, 1999). This fact implies that someone must be making the judgments. The research on judges' qualifications indicates that they must be knowledgeable about the content domain—judges who get items wrong set lower standards than those who answer them correctly (Chang, Dzuiban, and Hynes (1996)—but that they do not have to be experts on the content domain (Plake, Impara, and Potenza, 1994). ACT (1997) also recommends that the judges should represent a well-defined group or groups who have the necessary content knowledge and are familiar with students who are at the target grade level for the NAEP tests.

Methodology

The fourth structural component of the standard-setting process is the actual methodology for collecting information from judges about their recommendations for the standards. This

[2] For NAEP, individual examinees are not assigned to categories because there are no scores for individuals. Therefore, the Cohen, Kane, and Crooks (1999) definition cannot be directly applied. However, statistical procedures are used to estimate the number of examinees who would be assigned to each category if such assignments were possible.

component is the major focus of this section. A detailed discussion of the steps in a standard-setting process is provided in the next section of the paper. A variety of procedures will be described and discussed, with the goal of providing a general framework for evaluating potential methods for use with NAEP.

Reporting Mechanism

The fifth and final structural component for a standard-setting process is the method for reporting the results of the standard setting. Reporting for a standard-setting process usually includes cut-scores on the test score scale, and it may also include descriptions of behavior for persons who exceed the standard and examples of actual performance on test tasks. NAEP reports all of these types of information to help make the meaning of the achievement-level standards clear to the public (e.g., Williams et al., 1995).

Summary

Because standard-setting methods are often summarized using a single label (e.g., Angoff, benchmark), there is sometimes the impression that standard setting is a simple process. Standard setting is actually a very complex process that involves a number of components, each of which is critically important. The framing of initial policy is clearly important, and the translation of policy into meaningful content descriptions has been taking on added importance as more experience is gained with standard setting. For NAEP, content descriptions are now given formal approval before they are used to develop the standards for the achievement levels.

The importance of the qualifications of judges has received more attention in recent years. Judges need to be knowledgeable about the content covered by the test and the capabilities of the examinees. They also should represent a clearly defined population. Work by ACT has stressed the importance of the replicability of the standard-setting process, requiring that the selection of judges also be replicable.

The methodology used to collect judgments is clearly important. Without a sound methodology, the connection between policy, content, and cut-scores cannot be defended. Because of the importance of this component, many alternative standard-setting methods have been developed and substantial work is being done to evaluate the quality of alternative methods.

Finally, the way that standards are represented is of clear importance. Without clear communication, the value of the standards is lost. More research is needed to help identify methods for standard-based reporting.

Standard-Setting Methodology

Standard-setting methodology is the component of the standard-setting process that involves acquiring the judgments of individuals about the level of performance on the content domain needed to be considered above the standard. Acquiring the judgments involves four steps: (1) training, (2) collecting of judgments in one or more rounds, (3) providing supporting information

and feedback, and (4) estimating cut-scores. Each of these steps must be well designed and implemented if the standards that result are to be accurate and defensible.

Training

Training of judges has a number of goals. These include helping judges to understand (1) the policy that drives the standard setting, (2) the content framework that describes the content domain, (3) the mechanism that is used to collect their judgments, (4) the meaning of information that is provided to them as feedback on their performance, and (5) the interpretation of item characteristics and descriptions of student performance. The wealth of information that can be provided during implementation of the standard-setting methodology is quite large. Presenting that information clearly and concisely provides a challenge to the people involved in training. Extensive but confusing information can undermine the best methodology, resulting in excessive error in the judgments collected.

Training seems to be an underappreciated part of the standard-setting process. Most reports of standard-setting procedures provide little detail about training. Typically, two or three sentences are used to describe training methods. As a result, it is difficult to determine how well judges understand the tasks involved in setting the standards. A method may yield poor results because judges are poorly trained rather than because the method is flawed. More work needs to be done to determine the quality and type of training that is needed to support standard-setting methods.

Judgments

A major distinction among standard-setting methods is the type of judgment that is collected as part of the process. Some methods collect judgments about very fine-grained characteristics of test items, while others collect judgments about the performance of statistically defined groups of individuals. Previous catalogs of standard-setting methods have classified judgments into test-centered or examinee-centered categories (Jaeger, 1989). This distinction seems to be less useful as new procedures have been developed and as methods are applied to tests composed of extended tasks such as essays and science exercises. A better distinction among standard-setting methods may be the size or complexity of the unit that is judged.

At one extreme on a judgment task size or complexity continuum is the cognitive components model for standard setting proposed by McGinty and Neel (1996). One step in this method is the decomposition of items into the cognitive tasks that need to be accomplished for correct solution of the entire item. In a second step, judges indicate the probability that examinees who are above a standard should have of performing each cognitive task successfully. In preliminary studies, up to eight cognitive tasks were identified for each item. These cognitive tasks were at the level of detail of “translate words to numerals” or “apply basic addition facts” when the process was used on a third-grade mathematics tests.

At the other extreme of the judgment task size and complexity continuum is the cluster method proposed by Sireci, Robin, and Patelis (1997). This method requires the identification of clusters of examinees with similar profiles of content-level scores. The judgments required by the method are the number of clusters that represent unique profiles of content knowledge and the classification

of clusters of people according to the policy and content definitions for the standards. Cut-scores are set at points on the test score scale that best distinguish between adjacent clusters. Thus, this method requires judgments about the skills of groups of people, a very large and complex collection of information.

The size and complexity of the units that are the focus of the judgments can be placed along a continuum. Table 1 provides a graphic representation of that continuum. At the far left are standard-setting methods based on judgments related to fine details. Examples include the analysis of the specific skills needed to answer test items or detailed profiles of skills for individuals. At the far right of the continuum are standard-setting methods based on judgments of large collections of representations of content expertise. Examples include skills exhibited by collections of individuals or skills represented by individuals in extended bodies of work.

To clarify the meaning of this continuum, four different standard-setting methods have been placed along the task magnitude continuum as examples of variation in the task types that can be the focus of a standard-setting method. At one extreme is the cognitive components method suggested by McGinty and Neel (1996). Somewhat less detail oriented is the current ACT/ NAGB process that asks judges to indicate the probability of correct response that would indicate minimum acceptable performance on a test item. Further along the continuum is the Cohen, Kane, and Crooks (1999) generalized examinee-centered method. For this method, judges indicate how well full booklets of students' work match policy and content descriptions. Finally, the cluster approach of Sireci, Robin, and Patelis (1997) is placed at the far right because it focuses on the performance of groups of individuals rather than the work of a specific person.

Table 1. Continuum of Task Magnitudes for Judgment Tasks

Cognitive Components Model	ACT/NAGB Process	Generalized Examinee-Centered Method	Cluster Method
Judgments of Details			Judgments of Large Aggregates

The use of this continuum of judgment tasks is not meant to imply differences in quality of the methods, but only to show that methods vary quite dramatically on the types of judgments required. Later in this paper some criteria will be suggested for evaluating the promise of particular methods. The continuum of task magnitudes helps to identify likely problems with procedures, but any procedure, no matter where it is located on the continuum, might be a sound method if well designed and properly implemented. The evaluation of the methods will be concerned more with possible conceptual flaws than with the magnitude of the judgment tasks. Of course, the application of a method must be practical, and task magnitude may have some relationship to the practicality of the process.

Supporting Information and Feedback

Another way that standard-setting methods can be distinguished from one another is by the type and amount of information that is provided to the judges during the standard-setting process. Directly related to the amount of information is the number of rounds of judgments that are conducted. Generally, after information is provided to judges, they are allowed to use the information to guide revisions to the judgments they have made. If information is parceled out throughout a standard-setting process so that it can be absorbed and used more easily, then more rounds of judgments are used so judges are not overwhelmed by the quantity of information.

This section refers specifically to information designed to support the judgments made during the standard-setting task. The training component of the standard-setting process provides other information about policy, content, and the details of the judgmental process. Here the focus is on information about examinee performance and feedback on the outcomes of the judgment process. The types of information that are provided to support the judgment process can be arranged along a continuum, from types that tell judges how well they are performing the task to types that provide normative data about examinee performance. This continuum is shown graphically in Table 2.

Table 2. Continuum of Supporting Information Types

Rater Consistency	Rater Location	Consequences Data
Process Feedback		Normative Feedback

At the left end of the continuum is information that strictly deals with the functioning of judges in the process. The example given is a measure of rater consistency. Rater consistency tells judges if their ratings of test tasks are consistent within themselves. That is, have they rated hard tasks differently than easy tasks or good examples of student work differently than poor examples? From this type of feedback, judges can determine whether they understand what they are being asked to do and if they are applying the methods properly. In the current version of the ACT/ NAGB process, Reckase charts are used to provide this type of feedback to judges (ACT, 1998).

At the far right end of the continuum is information that deals strictly with the overall performance of examinees on the test or in test-related activities. If full distributions of scores are provided, these data are usually called “norms.” If norms are given early in the process, the standard-setting process is called normative or norm referenced. That is, judges take into account the number of persons who will be above a standard when they make their judgments. If little information is given about examinee performance and the focus is on the specific skills and content knowledge required, the standard-setting process is called “criterion referenced.”

The ACT/NAGB process is criterion referenced in the early rounds because mainly process feedback is given at that time, and it shifts toward more norm referenced, or normative, later in the process because consequences feedback is provided. Consequences feedback refers to

information about the percentage of students estimated to exceed each achievement level cut-score. Consequences feedback is provided as an example of normative information on the supporting information continuum.

Rater location information is placed at a point roughly in the middle of the continuum. This type of feedback tells judges how their standards relate to other judges. In a sense this feedback is normative because it tells the judges how their standards compare with a norm set by the entire group of judges. It is also process feedback because judges can tell if the ratings they provide result in a cut-score at the location on the score scale they intended. This example shows that information can be both normative and related to the functioning of the process.

There is not a strong connection between the standard-setting method used and the type of supporting information provided. A method based on task details can provide judges with process feedback and/or various types of normative information. A standard-setting method based on sorting students into categories can also supply process feedback and normative data. With rare exceptions, information types can be used with any type of standard-setting method. Because there is not a strong connection between supporting information and standard-setting method, it would be helpful if descriptions of standard-setting methods provided both a summary of the method and the types of information that are provided to guide the judges.

Cut-Score Estimation

After the judgments have been collected using a standard-setting method, the judgments must be aggregated and converted to a point (cut-score) on the reporting score scale for the test. This can be a highly technical process such as the maximum likelihood method used for the ACT/NAGB process (Davey, Fan, and Reckase, 1996) or a simple process such as computing the average of a set of ratings. A particular standard-setting method can use one or more procedures for converting judgments to cut-scores. Some of these methods may work very well, while others may result in inaccurate estimates of cut-scores. Because of the variety of cut-score estimation procedures that can be paired with a standard-setting method, the details of the way that cut-scores are determined need to be described along with the judgment method. It is certainly possible that a standard-setting method can yield questionable results because an inaccurate method was used for computing cut-scores even though the judges understood their task and gave well-considered judgments.

Summary

A standard-setting method is embedded within a larger standard-setting process that contains policy and reporting issues as well as the procedure for collecting judgments. The standard-setting method itself contains a number of steps. To understand how a particular implementation of a standard-setting method differs from another method, details must be provided about all the steps. For example, the ACT/NAGB process includes quite elaborate training for judges about NAEP, the role of NAGB, NAGB policy, the content frameworks, and the judgment process. The actual rating process consists of specifying the minimum probability of correct responses for dichotomously scored items and the mean response for polytomously scored items required to provide evidence that a person is in an achievement level category. This type of rating process

falls more toward the detailed end of the continuum of task magnitudes when multiple-choice items dominate a test and more toward the judgment of aggregates for the NAEP writing assessment.

The information provided to judges in the ACT/NAGB process ranges from process feedback to normative data and several levels in between. The information is both at the detailed item level and the whole booklet level. Judgments are converted to cut-scores using a sophisticated maximum likelihood estimation procedure. Details of this process are given in ACT (1997).

A somewhat different process is described in Cohen, Kane, and Crooks (1999) as the generalized examinee-centered method. This standard-setting process started with extensive training about policy, content, and process. The judges rated full booklets of student work on a 7-point scale, with score points defined by the policy statements. This magnitude of the task falls more toward the large aggregate end of the continuum. Judges received no information of the normative type, but did receive feedback on process and on rater location. Cut-scores were determined using linear equating methodology.

Step by step comparison between these two standard-setting methods indicates they are quite different on task magnitude, information, and cut-score estimation. They are similar in the nature and extent of the training given the judges.

Desirable Characteristics for a Standard-Setting Process

Because standard setting is a complex process that can be implemented in a number of different ways, it is difficult to identify firm criteria for determining whether a standard-setting procedure will likely give sound results. A very good standard-setting method can give poor results if the judges are not properly trained, if they do not have the necessary content background, or if the method is not implemented properly. Comparative studies can be misleading unless the methods being compared are implemented with equal levels of care. It is challenging, therefore, to come up with criteria for selecting a standard-setting method. Four criteria are presented here as at least reasonable characteristics for a good standard-setting method. Meeting these criteria will not guarantee that the method will work well, but not meeting the criteria likely indicates that the cut-scores produced using a method will be difficult to defend.

The four suggested criteria for a sound standard-setting methodology are: (1) minimal level of distortion in converting judgments to a standard, (2) moderate to low cognitive complexity of the tasks judges are asked to perform, (3) acceptable standard errors of estimate for the cut-scores, and (4) replicable process for conducting the standard setting study. Each of these criteria will be described, and they will be used to evaluate a variety of standard-setting methods in a later section of this paper.

Minimal Distortion of Judgments

There is no such thing as a true standard, but there is a theoretical cut-score that would be set by a judge if he or she totally understood the process, the test, the content, and the policy and had a true score on the test in mind as the standard. The question is whether the standard-setting

method can recover the theoretical cut-score assuming a judge performed every task consistently and without error. If the theoretical standard cannot be recovered for every possible value of the standard, then the method is flawed because it restricts or distorts the estimated values for the cut-scores.

A thought experiment can be used to check whether this criterion is met by a particular method. A specific score value can be assumed as the theoretical cut-score and the method can be analyzed to determine what a judge would have to do to have the method result in that cut-score. If there is no logical process that a judge can use to achieve that cut-score, the method has a serious problem.

An example of a method that has this type of flaw is the item score string estimation (ISSE) method that was pilot tested by ACT (Reckase and Bay, 1998). This method requires judges to indicate the item score that a person at the cut-score would most likely get for every item on a test. The response string generated by a judge is scored to obtain an estimate of the cut-score. Analysis of this procedure showed that if a judge performed the task perfectly, the cut-scores that were estimated would be more extreme than expected for many types of tests.

A simple example can show the problem. Suppose that a test is made up of 100 identical test items and the theoretical standard is 80% correct. This means that a person would have to answer items correctly 80% of the time to be exactly at the cut-score. For this test, the most likely response to each item is a correct response—the probability of an incorrect response would be 0.20. The result of following the instructions would be specifying a correct response for every test item as the most likely item score. The result would be a standard set at 100% correct rather than the theoretical cut-score of 80%. If the judge understood the method and applied it perfectly, it would be impossible to obtain the intended cut-score of 80% for this test. All possible values on the score scale should be considered as the theoretical cut-score to determine if the method distorts the judgments in a systematic way.

Moderate Cognitive Complexity

For a standard-setting method to be practical, the tasks that judges are asked to do must be within their capabilities. That is, they should not be asked to derive the Theory of Relativity unless they are well-trained theoretical physicists. The judges for standard-setting projects are often highly qualified individuals that have been selected because of their subject matter knowledge and experience with the examinee population. These individuals regularly perform challenging tasks in their everyday work. Still, within the context of a standard-setting meeting, they cannot be expected to perform extremely complex tasks with minimal training.

Evaluation of standard-setting methods on the criterion of cognitive complexity is very subjective. Without observing a standard-setting session, or getting feedback from judges, it is difficult to determine what is too cognitively complex. Experience with the various pilot studies and operational standard-setting studies for the ACT/NAGB process has shown that paper selection as a standard-setting method is extremely time consuming and fatiguing, and presses the limits of what judges can do (Bendixen, Price, and Webb, 1992). Also, several types of feedback related to consistency of ratings were too cognitively complex because they required

some knowledge of the principles of item response theory (American College Testing, 1993a). Other than these cases, judges consistently have indicated that they have no trouble doing the tasks asked of them in a standard-setting session. Yet, external critics of methods (e.g., Pellegrino, Jones, and Mitchell, 1999) suggest that a method based on specifying minimum item probabilities is too cognitively complex to yield meaningful results. The fact that the judgment tasks can be very challenging, and that some have suggested that they can be too challenging, suggests that procedures should be reviewed to determine the level of cognitive challenge involved. Those that have high cognitive loads should be evaluated carefully through pilot studies to determine if judges can perform the necessary tasks when given appropriate training.

Standard Error

All standard-setting methods yield cut-scores that contain error. Errors are because of differences in judges' interpretation of policy and content requirements, less than perfect judgments of the quality of work or the difficulty of items, limited samples of content domains, and inaccuracies in scoring the tests. Considering all the sources of error, the best that can be expected is to minimize the error in the estimation of the cut-score, rather than removing it altogether. The measure of error that is typically used for a cut-score is the standard error of estimation. This standard error is conceptually the standard deviation of the distribution of cut-scores that would be obtained if the entire standard-setting process were done multiple times with different judges and tests. The policy is assumed constant, as is the content domain.

The ACT/NAGB process has regularly reported standard errors that are fairly small (e.g., American College Testing, 1993b). The process that is used to compute the standard errors is to divide the judges into two groups and have each group work on different sets of test items. The comparison of the results from the two groups gives an estimate of the standard error.

Quality standard-setting methods should allow for estimating standard errors of cut-scores, and studies should be conducted to provide estimates of standard errors. Unfortunately, such studies are relatively rare, so this criterion can only be used to determine whether standard errors can be estimated in theory. Methods that do not even allow for the potential to estimate standard errors do not have enough theoretical grounding to be trustworthy.

Replicability

It should be possible to perform the same standard setting more than once in a way that would be expected to yield the same results. If this type of replication is possible, there can be some level of confidence in the cut-scores even if the replication is never carried out because of cost or other practical considerations. In this case, replication means that the process for selecting and training judges, the availability of test materials, etc., allow for an equivalent repetition of the entire process. For the ACT/NAGB process, a replicable sampling plan is used to select judges, and the training and analysis processes are well documented (ACT, 1997). Further, standard setting can be performed with subsamples of the NAEP item pools so independent replications are possible.

If a standard-setting method can only be performed by a carefully selected group of people, or under unique conditions, it is not replicable and cut-scores that result are in question.

Documentation for a standard-setting method should be complete enough to determine if replication is possible.

Summary

Four criteria have been suggested as minimum criteria for an acceptable standard-setting method. These are not extreme criteria. Most standard-setting methods should be able to meet them. They are applied in the next subsection of this section to a variety of methods, and the criteria do not exclude a great number of them. However, they do suggest ways that the methods can be improved, and they do identify some methods that have serious flaws. Certainly, other criteria can be suggested, but for the purposes of this section, these four criteria—minimal distortion of ratings, moderate cognitive complexity, acceptable standard error, and replicability—provide a framework for organizing the discussion of methods.

Review of Possible Standard-Setting Methods for NAEP

The review and evaluation of standard-setting methods that might improve on the current ACT/NAGB process is a difficult endeavor. Many of the procedures that have been suggested over the past decade have been used in very limited research studies, or merely described as possible procedures. On the other hand, the ACT/NAGB process has been applied numerous times and has been refined through a coordinated set of field tests and pilot studies. Comparing a new procedure to the current one is somewhat like comparing a movie from the early 1900's to a Rembrandt painting. The painting is the result of years of refinement in method and style of oil painting. The movie in the early 1900's represents new technology. So it is with many of the new standard-setting methods. They need to be evaluated on potential rather than on the results from the limited studies in the literature.

In the remainder of this section, a number of standard-setting methods are briefly described and the kernel of the method, the basic task performed by the judges, is evaluated using the four criteria listed above. These methods are also summarized in table 3. The methods are presented according to their location on the continuum of task magnitude described earlier, with those using ratings of details provided first. To be practically applied for NAEP, all these procedures would need extensive further development to connect the method to the policy and content frameworks, and to develop methods for reporting the results. None of the procedures are at the stage of development of the ACT/NAGB process, not even those used by State departments of education. Extensive work will need to be done before any of the methods can withstand the type of public scrutiny applied to the achievement levels.

Table 3. Summary of the Evaluations of Potential Standard-Setting Methods

Standard-Setting Method	Distortion of Ratings	Cognitive Complexity	Standard Error	Replicable
Cognitive Components	Yes, if number of components underestimated	High for identifying cognitive components	Higher than many	Yes
Modified Nedelsky	None	Moderate	Higher than many	Yes
Simple Angoff	Minor due to rounding to nearest 5%	Relatively low	Low	Yes
Item Mastery	Distortions due to mismatch between judges mastery probability and assumed mastery probability	Low for rating task; high for selecting mastery level and loss function	Might be large	Yes
Item Domain	Could be distortions caused by a mismatch between the mastery probability and judges' perceptions, and weaknesses in the item pool	Fairly high; need to match item characteristics to content descriptions	Likely large	Yes; need to replicate both mastery standard and item classification
Bookmark	Could be distortions caused by a mismatch between the mastery probability and judges' perceptions, and weaknesses in the item pool	Moderate	Small if items close together in difficulty near bookmark	Yes
Item Score Distribution	Little	High	Moderate	Yes
Holistic	Could be distortions caused by sample of papers available for classification	Moderate	Small	Yes
Anchor Based	Could be distortions if judges score papers differently than readers	Moderate	Small	Yes
Generalized Examinee Centered	Standards forced to be equally spaced by linear regression procedure	High	Small	Yes
Multistage Aggregation	Unknown	Moderate	Unknown	Yes
Contrasting Groups	Distortions with small samples	High	Small	Yes
Score Distribution	Cut-scores likely regressed toward the center of the score scale	High	Moderate	Yes
Cluster	Cut-scores limited by characteristics of student sample	High	Large	Yes

Cognitive Components Model

The cognitive components model (McGinty and Neel, 1996) has two phases to the kernel of the procedure. First, content experts analyze the test items to determine the cognitive components needed for their solution. For example, a simple addition item, 512 + 23 = , would be decomposed into (1) recognize "+" as a prompt for addition, (2) line numbers up vertically for addition, and (3) apply basic addition facts. Each item is decomposed into a number of the components. For the second phase, judges indicate the probability of correctly applying each cognitive component necessary to be minimally qualified according to the policy and content descriptions. Items have many cognitive components, and the probability a minimally qualified examinee will have of answering an item correctly is the product of probabilities for the components. The results are aggregated over test scores to determine a cut-score.

Recovering a hypothetical standard using this method requires that the two parts of the process work properly. First, the cognitive components that students use in responding to test items must be identified correctly. In particular, it is important that all major components be identified. If important components are missed, they are implicitly assumed to have a probability of 1.0 because no probability is estimated for them. The result is that the overall probability of correct response will be estimated as higher than it should be. The result would be setting a higher standard than if all the components were included. In the one study of this procedure, it was found to yield higher standards than the Angoff procedure, suggesting that some important components might have been missed. In theory, if all of the important components were identified, there would be no distortion of ratings in setting the standards.

The cognitive complexity of the rating task is no higher than other item-by-item procedures, but the cognitive complexity of identifying the cognitive components would seem to be quite high. It requires having a good understanding of the processes that students use to approach items and a clear understanding of what an item requires. This is not to say that the cognitive component analysis cannot be done, but it would seem that it must be done very carefully or the standards will have a positive bias.

In principle, the standard errors for this procedure can be computed, but it is important to check the amount of error induced by the identification of cognitive components. This would seem to be a major source of error. The fact that products of probabilities are used would also seem to emphasize errors. The method seems to be one that could be replicated as a check on the process.

This method is an interesting merger of cognitive science and psychometrics. The value of the procedure would seem to depend on how accurately the cognitive components of items can be identified.

Modified Nedelsky Method

Chang (1999) proposed a modified Nedelsky method for standard setting. When this method is applied, judges are asked to estimate the probability that an examinee who is above the standard will eliminate each wrong alternative from consideration. This is a modification to the standard Nedelsky method that asks judges to indicate which wrong alternatives an examinee will

eliminate (a dichotomous choice). The estimated *p*-value for an item is the sum of the probability estimates for choices plus 1.0, divided by the number of options. This method has not been implemented so no evaluative data are available.

Because there are no restrictions on the values judges can specify for the item options, it seems that judges could set any hypothetical cut-score without distortions. However, making judgments about probabilities for each of the response alternatives increases the cognitive complexity of the method over that of other item-by-item procedures. Judges will have to have a very good understanding of the knowledge and skills of examinees who are just above the standards.

This method can be replicated with different groups of judges and different sets of items. Standard errors can also be estimated if the judges are divided into two groups for an internal replication of the process. If judges have difficulty providing all of the probability estimates, the result may be fairly large standard errors for the cut-scores.

Simple Angoff Method

The Angoff method, as implemented in the current ACT/NAGB process (ACT, 1997), is quite an elaborate process that includes a variety of types of feedback and multiple rounds of ratings. As a basis for comparison for other methods, it is useful to include an unadorned Angoff procedure. Hurtz and Hertz (1999) describe the Angoff procedure as it is applied for setting licensure and certification standards in professional areas in the State of California.

For their implementation of the Angoff procedure, the kernel of the procedure involves having judges assign probability values to items from the range of 25% to 95% using increments of 5%. Judges initially provide ratings for the first few items, then discuss the results as a group. They then work through a set of items and have another discussion. The cycle of ratings of a set of items followed by discussion is performed two more times. Finally, the rest of the items on the test are rated. At the end there is a final discussion session followed by an opportunity to change previous ratings. The final ratings are used to set the cut-scores. The process they describe does not include multiple rounds of rating the same items or feedback on rater performance or examinee performance.

Hurtz and Hertz (1999) report estimates of standard errors from the results of multiple standard-setting studies on different content areas. They also estimate standard errors using the standard deviation of the standards set by individual judges. Clearly, the method allows estimation of standard errors. It is also replicable, either using an alternate form of a test or by splitting a panel of judges into two or more groups.

The cognitive complexity of this method does not seem particularly high. Hurtz and Hertz (1999) report that the method was used for eight different licensure examinations without undue difficulty. This is counter to the results presented in some studies, such as Impara and Plake (1998) that indicate that judges cannot accurately estimate the difficulty of test items. The difference in the standard-setting studies and studies like those reported by Impara and Plake (1998) are that for standard-setting studies, judges are selected carefully, and there is substantial training before rating test items. Also, there is usually a carefully developed content description

that guides the rating process. The Impara and Plake (1998) study used a random sample of teachers, did not include training, and did not provide judges with content descriptions. The differences in the way the studies were conducted may explain the differences in results.

Whether the judges can set a hypothetically selected standard using this version of the Angoff method is somewhat more difficult to determine. Because the judges were instructed to rate items using probability estimates that were in 5% increments, some cut-scores cannot be the result of the process. However, the difference between the theoretical standard and the one based on approximations from 5% increments will likely be fairly small.

Item Mastery Method

The item mastery method (Verhelst and Kaftandjieva, 1999) combines a number of features of other methods mentioned in this report—the ISSE method, the item domain method, and the bookmark method. This method is based on judges indicating whether students who are in a classification category "should be able to answer this item correctly." The responses are "yes" or "no." In this sense, the method is similar to the yes/no method suggested by Impara and Plake (1997) and the ISSE method. To estimate a cut-score using this method, a probability must be selected as a definition of mastery for an item. Verhelst and Kaftandjieva (1999) suggest having a second group of judges determine the mastery probability by indicating the percentage correct on a test overall that would indicate mastery. The procedure also requires the selection of a loss function for indicating the seriousness of placing the cut-score in the wrong place on the score scale. This information, plus estimated item response theory (IRT) item parameters for the items, is used to estimate the cut-scores.

This method has not been implemented for a formal standard setting. Therefore, it is difficult to determine how well it works in practice. A number of features of the model suggest that judges might not recover a hypothetical cut-score. First, if the mastery probability does not match the implicit probability that judges are using to determine what students should be able to do, there is likely to be a systematic difference between the hypothetical standard and the estimated cut-score. Furthermore, it is unclear how the loss function will interact with the ratings and the definition of mastery. A substantial amount of research will be needed before the operational characteristics of this procedure can be determined.

It does appear that the procedure is replicable and that the cognitive complexity of the item rating task is moderate. However, selecting a loss function and a mastery probability are high cognitive complexity tasks. The standard error of the estimates of the cut-scores could be quite large if the mastery probability corresponds to a region of the item response function that is relatively flat. In that case, a large number of item ratings will be needed to provide reasonably small standard errors.

Item Domain Method

Schulz, Kolen, and Nicewander (1999) suggest a method for standard setting that is based on the concept of item domains. This method requires two kinds of judgments. The first is an item-by-item judgment of the match of items on a test to the content definitions for a standard. Each item

is compared with the content descriptions for each level of performance to determine the content description that best matches the skills and content measured by the item. All the items are sorted into domains on the basis of the match to the content descriptions. If the content descriptions show a progression of improved performance, then the average percent correct for the items in each domain should show the reverse progression. That is, the items matched to the higher level content descriptions should be harder on average than those matched to the lower level content descriptions. If that is the case, then the standard-setting procedure can progress to the next step.

The estimation of cut-scores using this method is accomplished through the item response functions for each domain of items. The test characteristic curve is estimated for each set of items from the item characteristic curves. Next, the second judgment is made. A probability value is selected as the definition of mastery. Schulz, Kolen, and Nicewander (1999) selected 0.80 as the definition of mastery. Cut-scores are estimated by determining the value on the IRT score scale that yields the proportion correct for each domain. For NAGB achievement levels, domains would be defined for each achievement level. Items would be matched to the content descriptions for each achievement level. The test characteristic curves would be computed for each achievement level and the cut-scores would be determined using an agreed-upon probability value as a definition of mastery.

Replication of this procedure is possible, but it requires replication of both the classification of items according to content descriptions and the determination of the probability value that defines the requirements for mastery. The replication of classification of items to categories would seem to be fairly straightforward. The determination of the probability value for estimating the cut-scores would seem to be more difficult. Determining that value might be considered as part of policy.

Determining whether it is possible for judges to recover a hypothetical standard is difficult. The mechanism for setting a standard is the selection of items in the pool and matching them to content. To set a specific cut-score, judges would have to select a set of items that resulted in a test characteristic curve that gave the mastery proportion correct at the hypothesized cut-score. Doing this would seem to require careful matching of items to domains. If it were not possible to select items that gave the required form of the test characteristic curve, it would be impossible to set the hypothetical cut-score.

The cognitive complexity of this method would seem to be fairly high. Judges would have to determine whether an item required the skills and knowledge given in a content description. To do this accurately and consistently would seem to be very challenging. Inconsistencies in classification would likely result in large standard errors for cut-score estimates.

Bookmark Method

The bookmark method for standard setting is described on the web site for the Alaska Department of Education & Early Development (contact R. Smiley at: Richard_Smiley@eed.state.ak.us) and in a convention paper (Lewis et al., 1998). The method is being implemented for the State of Alaska by CTB/McGraw-Hill. This method asks judges to place “bookmarks” between the pages of books of items. The items are ordered in difficulty and

presented one item per page. The difficulty ordering is according to the point on the IRT-based score scale where students have a two-thirds probability of getting a correct response. Following placement of the bookmarks, the judges have a discussion session with the goal of reaching consensus on their placement. If consensus is not reached, the average bookmark placement is used for the cut-score.

A concern that has been raised about the bookmark procedure is whether the standards that are set with the method depend on the probability value used to order the items. When different values are used, the ordering of the items changes somewhat. Whether judges can set a hypothetical standard is related to the amount of variation in item order for different probability values. If the judges understand the relationship between the probability of correct response and the item placement, they can set the standard at least close to a hypothetical standard. This is only true if there are items close together on either side of the intended standard. However, if the judges have a different probability in mind than the one used for ordering the items when they place their bookmark, the cut-score estimate will likely be higher or lower than the hypothetical cut-score.

The cognitive complexity for this method seems moderate because the judges need to only decide the two items that are on either side of their standard, or the group of items that are near the standard, and place the bookmark in the middle of them. The standard errors for the procedure are likely to be small if a substantial number of items are near the bookmark. However, the standard error could be large if the bookmark is placed where items have substantial variability, or if there is a mismatch between judges' perceptions and the preset definition of mastery. This method should be replicable.

Item Score Distribution Method

The *Stanford Achievement Test Series, Ninth Edition* reports results using performance standards that are similar to NAEP achievement levels. These standards were set using methods very much like the ACT/NAGB process. Judgments for the performance items on the test were made using the score distribution method described by Luecht (1993). This method asks judges to "assign percentages of each borderline group that should receive each score points, with 3 interpreted as essentially correct, 2 as partially correct, 1 as marginally acceptable, and 0 as essentially incorrect" (Harcourt Brace, 1997).

This method applies to moderately sized tasks such as short essays. With a good understanding of the methodology, judges should be able to produce a hypothetical standard because a wide variety of distributions can be specified during the rating process. Cognitive complexity is the challenge to this procedure. Judges must not only indicate the typical level of performance that is expected by a minimally competent examinee, but they must also indicate how much the performance will vary across examinees who meet the requirements for being minimally competent. No data are available on how well judges can do the variance estimation task.

The item score distribution method should allow for the computation of standard errors and, in theory, the process is replicable. Overall, this would seem to be a viable method, with the only concern being with the cognitive complexity of the judgment task. The level of confidence in this

method will depend on evaluations of how well judges can specify probability distributions across the categories of a scoring rubric.

Holistic Procedure

Jaeger and Mills (1997) proposed a procedure that has a number of similarities to the anchor-based procedure suggested by Hambleton and Plake (1996). Both of these procedures develop a rating scale for classifying student work that is directly connected to the performance descriptions for the standards. For each level of the standards, panelists are asked to classify student work as being low, medium, or high examples of that category of performance. The major difference between the holistic procedure and the anchor-based procedure is that the holistic procedure asks panelists to consider parts of the student's body of work first, and then to rate the entire body of work, rather than working with full test booklets.

The study that was performed to evaluate the holistic procedure used test booklets that had both multiple-choice and performance assessment activities. Panelists first classified the performance on the multiple-choice portion of the test into 1 of 12 categories for a selection of student papers. The 12 categories were defined as high, medium, and low performance in each of 4 performance levels—Advanced, Proficient, Apprentice, and Novice. Panelists then classified the student work on the performance assessment exercises into the same 12 categories. Finally, they classified the full test booklets into the performance categories. In all cases, panelists were given the scores for the portion of the test that they were classifying.

Standards were set using both the weighted mean of performance on either side of the performance-level boundary—the weighted mean of the high-category papers from the lower category and the low-category papers from the higher category. Linear and cubic regression procedures were also considered to smooth the relationship of the cut-scores. The authors preferred the cubic regression procedure.

To set a hypothetical standard, judges have to select a set of papers that will provide the desired weighted mean. This would appear to be a challenging activity. Furthermore, the actual weighted mean would depend on the number of papers in the sample that would be classified into each category. For example, suppose that the panelist was selecting papers for the Advanced/Proficient cut-score. If they found many papers for the high end of the Proficient category, but hardly any for the low end of the Advanced category, the standard would be set low. If the opposite situation occurred, the standard would be set high. Thus, the set of papers available for classification affects the accuracy of estimation of the standards.

The classification task would seem to be of moderate cognitive complexity, but of higher complexity than classification of individual pieces of work. The task seems replicable and the research reported on this method reported fairly small standard errors for the cut-scores estimated from the method. Overall, the method seems to have potential, but it has only been tried in one limited study. More work would need to be done to refine the procedure.

Anchor-Based Procedure

Hambleton and Plake (1996) have developed a standard-setting method for use with performance assessment tasks that asks judges to classify examples of student work into categories along a 12-point scale that is directly connected to the performance standards. Each performance category (e.g., Proficient) is divided into low, middle, and high subcategories. For the NAEP achievement levels, there would be 12 categories in all if the Below Basic category were also divided into 3 levels. Judges are given 50 student papers and are asked to sort them into the 12 categories. Cut-scores on the tasks are set by averaging the scores on the papers that are classified in the top level of one category and the lowest level of the next higher category. They also suggest fitting a curve to the relationship between the sequential numbers for the categories and the scores on the test.

It would seem that judges could set a hypothetical standard if the scores on the papers put into adjacent categories were not too different and if the judges knew the scores assigned to the papers. Then they could select papers that yielded the mean that corresponded to the hypothetical standard. However, if the judges apply the scoring rubrics differently than the way the exercises were actually scored, the results would be a shift in the cut-score. Hambleton and Plake (1996) did not give the judges the scores on the papers for their study, but they suggest that supplying scores might be a good idea. A poor sampling of papers might also make it impossible to select papers that yielded the required mean.

The cognitive complexity of this task seems moderate, although sorting papers into 12 categories is beyond what is typically done when scoring performance assessments. Most scoring rubrics have three to six categories. Classifying papers into 12 levels is probably more difficult than traditional performance assessment scoring. This method allows the computation of standard errors using subgroups of judges and papers, and the process is replicable if judges and papers are selected carefully.

Generalized Examinee-Centered Method

The generalized-examinee centered method was implemented by Cohen, Kane, and Crooks (1999) for setting three cut-scores on the tests from a State assessment program. Cut-scores were determined between the following four performance categories: Minimal, Partially Proficient, Proficient, and Advanced. The method was applied to full test booklets, a fairly large task component on the task magnitude continuum. After extensive training, judges were asked to rate booklets using a 7-point rating scale directly connected to the performance-level definitions. Booklets classified in the Minimal category were given a rating of 1, Partially Proficient a 3, Proficient a 5, and Advanced a 7. The borderlines between the categories were assigned 2, 4, and 6 respectively. To obtain cut-score estimates, the rating scale was equated linearly to the test score scale and the scores that corresponded to 2, 4, and 6 were the estimates of the cut-scores.

Whether judges could set a hypothetical cut-score is an interesting technical question. Because linear equating was used, and because 2, 4, and 6 are equally spaced on the rating scale, the cut-scores on the test score scale must also be equally spaced. This means that judges did not have full freedom in setting cut-scores as the method was implemented for this study. This restriction

could be removed by using a nonlinear equating method, but the developers indicated "that linear functions fit the data as well as any of the nonlinear functions we tried." The rating process may not be accurate enough to allow nonlinear functions to be estimated.

A related issue is the cognitive complexity of the task that the judges are asked to do. They are asked to evaluate full booklets of students' work relative to the performance standards. This would seem to be a difficult task. They do not score the booklets, and a major point is made about the distinction between scoring and rating. Although the judges received extensive training, in some cases the ratings correlated only moderately with the total scores on the tests.

Standard errors were estimated for the results of the method, and the method is replicable. Overall, this method seems to have some technical problems that need to be fixed before it can be given serious consideration. If the restriction on the spacing of the cut-scores can be overcome, and if there is evidence that the judges can meaningfully evaluate full booklets of work without scoring the booklet or receiving scores, then the procedure may have promise. It does include the innovation of fitting a function to the rating data to stabilize the results. This is both a power of the method and the cause of one of the problems.

Multistage Aggregation Method

De Champlain et al. (1998) have pilot tested a standard-setting method in the medical patient management problem context that asks judges to rate increasing amounts of performance information over rounds. The judges also provided information about how ratings should be combined to estimate a cut-score. On the continuum of task magnitude, this method starts in the middle with ratings of large items (e.g., a particular patient management problem) and progresses to sets of problems, and finally to overall performance. The method covers quite a large range at the right side of the task magnitude continuum.

At each round of rating, the judges are asked to use a 5-point rating scale that has 3 as just adequate performance, 5 as clearly adequate, and 1 as clearly inadequate. In the first round, this rating scale is used for a 25-item, dichotomously scored checklist. This would be the equivalent of having judges rate specific sets of content-related items on NAEP. In the second round, the rating scale is used to report judgments of competency on a specific problem based on the profile of performance on multiple dimensions. At a third round the rating scale is used to rate profiles of performance based on responses to multiple problems. Finally, a dichotomous pass/fail decision is made using all the information. The method the raters used for combining the information was determined using logistic regression to predict the pass/fail decision from the ratings of each problem. This analysis allowed investigations of the compensatory or noncompensatory nature of the overall judgments.

This method was presented in the context of research on ways for combining multiple ratings, so it is difficult to determine what the details of a refined rating process would be. The study did not result in a cut-score estimate, although it would be easy to imagine a cut-score estimation procedure that is either the sum of the scores for a set of problems or a count of the number of problems that were handled in a satisfactory way. Because of the incomplete development of this

rocedure at this time, it is not possible to determine if judges could recover a hypothetical cut-score.

The cognitive complexity of this method seems moderate because the judgments of the larger units of performance build on the judgments of smaller units of performance. Whether standard errors of the cut-scores can be determined will have to wait until a formal score scale is specified because the standard errors are directly related to the scale. It would seem that this method could be replicated if the selection of judges is replicable.

Contrasting Groups Method

Although the contrasting group method is not a newly developed standard-setting method, it is included here for comparison purposes because it is frequently suggested as an alternative to the modified Angoff method. Under the current NAEP testing model that does not produce student-level scores, it would not be possible to implement a contrasting groups-type approach. If the NAEP testing model were changed to give student-level scores, this method could be given more serious consideration.

The contrasting groups method requires that judges classify individuals into categories according to the policy and content descriptions for the categories. These categorizations are made without knowledge of the scores on the test. Rather, other information about persons' capabilities, either through personal contact or through review of other performance data, is used to inform the classifications. After the classifications are made, the test score distributions are estimated for each classification category. The differences in the score distributions are used to determine cut-scores for the standards. Numerous methods have been proposed for estimating the cut-scores. Different cut-score estimation methods yield somewhat different results.

The task magnitude for this method is fairly large. Judges need to understand the capabilities of an individual relative to the policy and content descriptions without the benefit of scores on the test that is the focus of the standard setting. This could require digesting substantial amounts of observational data, or reviewing performance data in a variety of forms.

Cizek and Husband (1997) performed a simulation study that directly addressed the issue of whether judges could recover a hypothetical cut-score. They used a cut-score estimation method based on the point that score distributions from different categories intersected. They also applied a smoothing procedure to the distributions to remove irregularities due to small sample sizes. They found that a sample size of 24 was too small to yield stable results, but that sample sizes of 102 and 1,020 did allow judges to recover a hypothetical cut-score fairly well.

The cognitive complexity of the contrasting groups method is fairly high. Judges need to consider a complex set of skills and knowledge about an individual without access to a summary of skills. This information needs to be compared to the policy and content descriptions to determine the appropriate classification of an individual. This would seem to be quite a challenging task.

The contrasting groups method would seem to be replicated easily. Either separate groups of judges could do the classifications or the same group of judges could classify random samples from the same population. The standards established from the set of replications can be compared to get an estimate of the error in the process. Other than the Cizek and Husband (1997) study, no literature on the standard error of estimates of cut-scores for the contrasting groups method has been found.

Score Distribution Method

On the far right of the task magnitude continuum is a method that asks judges to estimate the full distribution of scores for examinees who just barely exceed the criteria specified by the policy definitions and the content descriptions (Poggio and Glasnapp, 1994). More specifically, the judges are asked to respond to the following question: "What should be required as the *minimum . . . acceptable performance distribution* for a group of 100 regular education students to be judged as performing at each Proficiency Scale level on the total set of items?" The judges are asked to put their estimated score distribution on a percentage correct scale for the entire test. This task is performed after reviewing each test item and training on the policy and content definitions for the proficiency standards. The cut-score for each proficiency standard is estimated by computing the mean or median of the estimated distribution for each judge, then averaging the estimates across judges.

Whether judges can recover a hypothetical cut-score using this procedure is an interesting one. Initially it would seem that judges could set the cut-score anywhere on the score scale by careful specification of a score distribution. For cut-scores in the middle range of score values, this would be fairly easy to do, because a symmetric distribution around a point would yield the mean at the middle of the distribution. However, as the hypothetical cut-score becomes closer to the highest or lowest score on the test, producing a distribution with the desired mean or median becomes more challenging. The distributions would have to be either very skewed or have very little range. For example, to get a mean score near the maximum possible on the test, the distribution would have to be very tight around the desired cut-score. This analysis suggests that this method likely has a statistical bias such that cut-score estimates would be closer to the center of the score range than hypothetical cut-scores. The effect would be greatest at the extremes and smaller in the middle ranges of the score scale. It may be that extensive training on the relationship between the mean and the shape of score distributions could reduce or remove this bias.

The cognitive complexity of this standard-setting method would seem to be fairly high. Judges need to study all the items on the test, then generate a full score distribution that describes the expected performance of minimally qualified examinees. In the one study of this procedure that was found in the literature, the judges did not seem to have any more trouble with this method than with the modified Angoff method, but the characteristics of the judges or their training were not described in the study.

It would seem that this method could be replicated using multiple sets of judges and parallel test forms and that it would be possible to estimate a standard error for the cut-score estimates.

Overall, the greatest concern about this procedure would seem to be with the cognitive complexity of the task, and the possible statistical bias in cut-score estimates.

Cluster Method

Another method at the far right of the task magnitude continuum is the cluster method proposed by Sireci, Robin, and Patelis (1997). For this method, the performance on a test is initially reported subscores on each content area. These subscores are used for the purpose of analysis only rather than for formal reporting. A cluster analysis of student performance is performed using a hierarchical method to explore the number of clusters, then a K-means cluster analysis is run after the number of clusters had been determined. The logic of the procedure is that students within clusters are highly similar and that cut-scores should fall between clusters of similar students. The judgment part of the process is to decide on the appropriate number of clusters and the connection between the performance of a cluster of students and the content and policy descriptions of standards. In this study, psychometric staff made these judgments.

The issue of whether judges can set a hypothetical standard is very complex for this method. There is an assumption that standards should only be set where there are naturally occurring gaps in performance. If the gaps do not occur at the location of a hypothetical cut-score, the procedure could not set it there. This is particularly a problem because in the study of this procedure, the clustering of examinees did not replicate across random samples of examinees from the same population. One sample had natural breaks at six clusters, while the other sample had breaks at four clusters.

Although the idea of matching clusters of examinees to the performance standards rather than individuals or item performance is an interesting one, it seems this method does not meet either the replication or the match to hypothetical standard criteria. These problem areas also imply that the standard errors for the cut-scores would be quite large. Moreover, the cognitive complexity would seem to be high because the focus is on the comparison of large groups of persons to the policy and content descriptions.

Summary and Conclusions

This section presents a general framework for describing the components of standard-setting procedures, develops some criteria for evaluating standard setting procedures, and then describes a number of potential methods. Quite a bit of creative work has been done in the area of standard setting over the past few years, as is shown by the variety of standard-setting methods that have been suggested. They vary from rating with the detailed cognitive components hypothesized to exist in test items to classifying large clusters of examinees.

Some new methods have not been included in this report because they either could not be adapted easily for use with NAEP, or because they were already very similar to procedures that have been used for NAEP achievement levels setting. Noncompensatory methods such as dominant profile method (Plake, Hambleton, and Jaeger, 1997) and the extended Angoff procedure (Hambleton and Plake, 1995) are examples of such methods. The Plake, Hambleton, and Jaeger (1997) method combines subscores in a ways that is inconsistent with how NAEP is

scored, and the Hambleton and Plake (1995) method is very similar to the method already used for setting standards on performance items for NAEP. This is not to suggest that there is any problem with the methods that have not been included. It only means that major changes would have to be made to NAEP before some of these methods could be applied.

Although many methods have promising features, in general the methods have not been evaluated thoroughly and none have been subjected to the type of scrutiny given the ACT/NAGB process. This implies that substantial developmental work will be needed before any of the methods can be used for a program like NAEP.

Most of the methods described here have been used in limited research or pilot studies. They have not been used in the full context of a formal standard-setting study. The way that policy and content information is presented to judges needs to be determined. The types of information and feedback that will be provided to judges needs to be designed. The number of rounds of rating required to stabilize results and possibly reach consensus needs to be determined. Combinations of these methods may yield the best overall process. As in the multistage aggregation method, it might be useful to use one method for a first round of judgments, and another more holistic method for later rounds of judgments. The ACT/NAGB process has been moving in that direction.

Of the methods presented here, the bookmark method, the anchor-based procedure, the generalized examinee-centered method, and the multistage aggregation method seem to have the most promise. This mild endorsement must be qualified by a caution that the final form for any of these methods is unknown. Thorough study of any of them may identify serious shortcomings. Several criteria were suggested for evaluating standard-setting methods in this paper. As these methods receive further development, it is hoped that these criteria will help guide the refinement and improvement of the methods.

References

ACT, Inc. (1998). *Briefing booklet: 1998 writing NAEP achievement levels-setting. Iowa City, IA.*

ACT, Inc. (1997). *Developing achievement levels on the 1998 NAEP in civics and writing: Design document.* Iowa City, IA.

American College Testing (1993a). *Description of mathematics achievement levels-setting process and proposed achievement level descriptions (Volume 1).* Iowa City, IA.

American College Testing (1993b). *Setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading, and writing: A technical report on reliability and validity.* Iowa City, IA.

Bendixen, A., Price, B., and Webb, M.W. (1992). *Setting achievement levels for the 1992 NAEP writing assessment.* Paper presented at the annual convention of the National Council of Teachers of English.

Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education,* 12:151–165.

Chang, L., Dziuban, C.D., and Hynes, M.C. (1996). Does a standard reflect minimal competency of examinees or judge competency? *Applied Measurement in Education,* 9:161–173.

Cizek, G.J. and Husband, T.H. (1997). *A Monte Carlo investigation of the contrasting groups standard setting method.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Cohen, A.S., Kane, M.T. and Crooks, T.J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education,* 14:343–366.

Davey, T., Fan, M., and Reckase, M.D. (1996). *Some new methods for mapping ratings to the NAEP θ-scale to support estimation of NAEP achievement level boundaries.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

De Champlain, A.F., Margolis, M.J., Ross, L.P., Macmillan, M.K., and Klass, D.J. (1998). *Setting test-level standards for a performance assessment of physicians' clinical skills: A process investigation.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

G. & C. Merriam Company. (1977). *Webster's New Collegiate Dictionary.* Springfield, MA.

Hambleton, R.K. and Plake, B.S. (1995). Using an extended Angoff procedure to set standards on complex assessments. *Applied Measurement in Education,* 8:41–56.

Hambleton, R.K. and Plake, B.S. (1996). *An anchor-based procedure for setting standards on performance assessments.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Harcourt Brace & Company (1997). Content and performance standards. *Stanford 9 Special Report.* San Antonio.

Hurtz, G.M. and Hertz, N.R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? *Educational and Psychological Measurement,* 59:885–897.

Impara, J.C. and Plake, B.S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement,* 34:353–366.

Impara, J.C. and Plake, B.S. (1998). Teacher's ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement,* 35:69–81.

Jaeger, R.M. (1989). Certification of student competence. In: R. L. Linn (Ed.). *Educational measurement (3rd edition).* New York: American Council on Education and Macmillan.

Jaeger, R.M. and Mills, C.N. (1997). *A holistic procedure for setting performance standards on complex large-scale assessments.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Kahl, S.R., Crockett, T.J., DePascale, C.A., and Rindfleisch, S. L. (1995). *Setting standards for performance levels using the student-based constructed-response method.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Lewis, D.M., Green, D.R., Mitzel, H.C., Baum, K., and Patz, R.J. (1998). *The bookmark standard setting procedure: Methodology and recent implementations.* Paper presented at the 1998 National Council for Measurement in Education Annual Meeting, San Diego.

Luecht, R.M. (1993). *Using IRT to improve the standard setting process for dichotomous and polytomous items.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta.

McGinty, D. and Neel, J.H. (1996). *Judgmental standard setting using a cognitive components model.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

National Assessment Governing Board (1990). *Setting appropriate achievement levels for the National Assessment of Educational Progress: Policy framework and technical procedures.* Washington, DC.

National Assessment Governing Board (1998). *Civics framework for the 1998 National Assessment of Educational Progress.* Washington, DC.

Pellegrino, J.W., Jones, L.R., and Mitchell, K.J. (1999). *Grading the Nation's report card.* Washington, DC: National Academy Press.

Plake, B.S., Hambleton, R.K., and Jaeger, R.M. (1997). A new standard-setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement,* 57:400–411.

Plake, B.S., Impara, J.C., and Potenza, M.T. (1994). Content specificity of expert judgments in a standard-setting study. *Journal of Educational Measurement,* 31:339–347.

Poggio, J.P. and Glasnapp, D.R. (1994). *A method for setting multi-level performance standards on objective or constructed response tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Reckase, M.D. and Bay, L. (1998). *Analysis of methods for collecting test-based judgments.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Diego.

Schulz, E.M., Kolen, M.J., and Nicewander, W.A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement,* 23:347–362.

Sireci, S.G., Robin, F., and Patelis, T. (1997). *Using cluster analysis to facilitate the standard setting process.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Verhelst, N.D., and Kaftandjieva, F. (1999). *A rational method to determine cutoff scores (Research Report 99–07).* Enschede, The Netherlands: University of Twente, Faculty of Educational Science and Technology, Department of Educational Measurement and Data Analysis.

Williams, P.L., Lazer, S., Reese, C.M., and Carr, P. (1995). *NAEP 1994 U.S. history: A first look.* Washington, DC: U.S. Department of Education.

SECTION 4

A Description of the Standard-Setting Procedures Used by Three Standardized Achievement Test Publishers

Robert A. Forsyth University of Iowa

November 2000

A Description of the Standard-Setting Procedures Used by Three Standardized Achievement Test Publishers

Robert A. Forsyth

Since the adoption of standards-based reporting of achievement test results by the National Assessment of Educational Progress (NAEP) in the early 1990s, this type of reporting has become commonplace. Most State testing programs provide standards-based reports, and publishers of standardized achievement test batteries also include standards-based reports as part of their services. In this section, the procedures used to establish performance standards for three of the most widely administered elementary school standardized achievement batteries will be described. These three batteries and their publishers are:

1. *TerraNova* (CTB/McGraw-Hill, 1997a).

2. *Stanford Achievement Test Series* (Harcourt Brace, 1996a).[1]

3. *Iowa Tests of Basic Skills* (Riverside Publishing, Hoover, et al., 1996a).[2]

This section is divided into three major subsections:

1. Frameworks and Achievement Levels.

2. Methods Used To Derive Cut-Scores.

3. Outcomes of the Standard-Setting Process.

Within each of these subsections, the materials and procedures employed by the three publishers are discussed. These publishers have established performance standards in several content domains; however, in this section all illustrations and data in the main body of the text are from the reading area.

The primary purpose of this section is to describe the standard-setting procedures of three test publishers, not to evaluate the relative merits of those procedures. Likewise, this section does not evaluate the merits of the publishers' procedures relative to the NAEP standard-setting

[1] At the time of publication of *Stanford Achievement Test Series*, the official name of the publisher was Harcourt Brace Educational Measurement. Recently, this name was changed to Harcourt Educational Measurement. Because most of the sources of information about the *Stanford* were published before the name change, Harcourt Brace will be used to identify this publisher throughout the paper.

[2] It should be noted that I am a coauthor of another standardized achievement test battery published by Riverside Publishing. This battery, the *Iowa Tests of Educational Development,* is intended for students in grades 9 through 12.

procedures.[3] However, as a point of reference, some comparisons between the NAEP and the three publishers' procedures are provided.

Frameworks and Achievement Levels

Frameworks

Assessments typically begin with the development of a framework that defines the domain for the assessment. The 1998 NAEP Civics Assessment, for example, began with the development of the Civics Framework (National Assessment Governing Board, 1998). Likewise, publishers of standardized achievement tests begin with a set of domain specifications for their tests. For a given domain, these specifications usually include both a cognitive (process) dimension and a content dimension.

For the three achievement batteries considered in this section, descriptions of these test specifications can be found in the following publications:

1. *Teacher's Guide to TerraNova* (CTB/McGraw-Hill, 1999).

2. *Stanford Achievement Test Series: Compendium of Instructional Objectives* (Harcourt Brace, 1996b).

3. *Iowa Tests of Basic Skills: Interpretive Guide for Teachers and Counselors, Form M* (Hoover et al., 1996b).

As an example of such specifications, table 1 shows the content and process categories that form the basis of the *Stanford* reading comprehension tests.

[3] Actually, it is unreasonable to refer to "the NAEP standard-setting procedures," because NAEP procedures have changed over the years. For example, in the early NAEP assessments that reported performance using achievement levels (e.g., mathematics and reading in 1992), standard-setting panelists created content-specific achievement level descriptions after the performance levels (cut-scores) were set; in the latest assessments (civics and writing), content-specific achievement level descriptions were developed before the standard-setting panelists were convened (Hanick and Loomis, 1999).

Table 1. Major Content and Process Categories for the *Stanford* Reading Comprehension Tests[a]

Content Categories

Recreational
Demonstrate the ability to construct meaning with material typically read for enjoyment.

Textual
Demonstrate the ability to construct meaning with material typically found in grade-appropriate textbooks and other sources of information.

Functional
Demonstrate the ability to construct meaning with material typically encountered in everyday life situations.

Process Categories

Initial Understanding
Demonstrate the ability to comprehend explicitly stated relationships in a variety of reading selections.

Interpretation
Demonstrate the ability to form an interpretation of a variety of reading selections based on explicit and implicit information in the selections.

Critical Analysis
Demonstrate the ability to synthesize and evaluate explicit and implicit information in a variety of reading selections.

Strategies
Demonstrate the ability to recognize and apply text factors and reading strategies in a variety of reading selections.

[a]Adapted from *Stanford Achievement Test Series, Ninth Edition: Compendium of Instructional Objectives* (Harcourt Brace, 1996b). These categories are used for all reading test levels.

Achievement Levels

As shown in table 2, two of the publishers (Harcourt Brace and Riverside) decided to use the same number of achievement levels or performance categories (four) as NAEP and also to use NAEP labels for these levels. CTB/McGraw-Hill uses five performance categories and gives the highest two categories the same label as the highest two levels of NAEP.

Table 2. Labels for Achievement Levels

CTB/McGraw-Hill	Harcourt Brace	Riverside	NAEP
Advanced	Level 4, Advanced	Advanced	Advanced
Proficient	Level 3, Proficient	Proficient	Proficient
Nearing Proficiency	Level 2, Basic	Basic	Basic
Progressing	Level 1, Below Basic	Below Basic	Below Basic
Step 1[a]			

[a] At grades 1 and 2, this level is labeled "Starting Out."

In the NAEP standard-setting process, policy definitions that were established by the National Assessment Governing Board are provided for the Advanced, Proficient, and Basic achievement levels. These definitions, as stated for the 1998 NAEP Civics Assessment, are given below:

- *Advanced:* This level signifies superior performance.
- *Proficient:* This level represents solid academic performance and competency over challenging subject matter.
- *Basic:* This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for, proficient work in grades 4, 8, and 12 (National Assessment Governing Board, 1998, p. 41).[4]

Of the three publishers, only Harcourt Brace provided "NAEP-policylike" definitions as part of its standard-setting process. In fact, the Harcourt Brace definitions, as shown below, are very similar to the NAEP definitions.

- *Advanced:* Signifies *superior performance* beyond grade-level mastery.
- *Proficient:* Represents *solid academic performance,* indicating that students are prepared for the next grade.
- *Basic:* Denotes *partial mastery* of the knowledge and skills that are fundamental for satisfactory work.
- *Below Basic:* Indicates *less than partial mastery*. (Harcourt Brace, 1997c, p. 41)

As the above statements illustrate, policy definitions are very generic and can be applied to a variety of content domains. When such definitions are provided, they are usually translated into descriptions of what students who have achieved a specific achievement level know and can do

[4] These policy definitions have been used since the 1994 NAEP assessments. (See, for example, Campbell et al., 1996.) The original policy definitions used with the 1990 and 1992 NAEP assessments can be found on page 5 of Bourque and Garrison (1991).

in a given content domain. These specific descriptions are often referred to as Achievement Level Descriptions (ALDs). In some standard-setting situations, ALDs are developed before implementation of the procedures for setting performance levels or cut-scores. These initial ALDs may be revised after the cut-scores have been established. In the Harcourt Brace standard-setting process, the policy definitions noted above were translated into test-specific achievement level descriptions during an orientation session. At this session, "teachers were provided with descriptions of the performance levels [policy definitions], and guided in constructing opera-tional definitions of the Performance Standards [achievement levels]" (Harcourt Brace, 1997a, p. 3).

These operational definitions were used by the standard-setting panelists to help them set the cut-scores. These operational definitions could not be located, and it appears that the only ALDs published by Harcourt Brace are the policylike definitions. However, Harcourt Brace does provide additional information about what students at each performance level can do. This additional information is discussed in more detail in the last subsection.

Neither CTB/McGraw-Hill nor Riverside provided policylike definitions for their achievement levels. However, both publishers did develop ALDs before beginning the process of setting cut-scores. CTB/McGraw-Hill, much like Harcourt Brace, had standard-setting panelists create operational definitions as described below:

> The first task required of committee members [panelists] was to carefully review the ordered-item booklets [booklets with items ordered by difficulty] and to consider the knowledge and skills required to respond to the items as they increased in difficulty. Participants used their own expectations and the ordered-item booklets to operationalize their expectations in terms of actual test content to obtain initial performance level descriptors that reflected what students should know and be able to perform as measured by the test and not relative to an idealized curriculum. After extended discussion, the committee members assembled the initial performance level descriptors.[5]
> (CTB/McGraw-Hill, 1997b, pp. 237–38)

These operational ALDs were modified by the panelists after the final cut-scores were established. The final descriptions for each grade group and content area are reported in Part 4 of the *TerraNova Performance Levels Handbook* (CTB/McGraw-Hill, 1997c). Table 3 provides an example of the final ALDs developed by the reading panel for grades 3, 4, and 5. Procedures used to derive these final descriptions will be discussed in the last subsection.

[5] These "initial performance level descriptors" could not be located.

Table 3. CTB/McGraw-Hill Final Achievement Level Descriptors for Reading: Grades 3–5[a]

Advanced	Students use analogies to generalize. They identify a paraphrase of concepts or ideas in texts. They can indicate thought processes that led them to a previous answer. In written responses, they demonstrate understanding of an implied theme, assess intent of passage information, and provide justification as well as support for their answers.
Proficient	Students interpret figures of speech. They recognize paraphrases of text information and retrieve information to complete forms. In more complex texts, they identify themes, main ideas, or author purpose/point of view. They analyze and apply information in graphic and text form, make reasonable generalizations, and draw conclusions. In written responses, they can identify key elements from text.
Nearing Proficiency	Students use context clues and structural analysis to determine word meaning. They recognize homonyms and antonyms in grade-level text. They identify important details, sequence, cause-and-effect, and lessons embedded in the text. They interpret characters' feelings and apply information to new situations. In written responses, they can express an opinion and support it.
Progressing	Students identify synonyms for grade-level words and use context clues to define common words. They make simple inferences and predictions based on the text. They identify characters' feelings. They can transfer information from text to graphic form, or from graphic to text form. In written responses, they can provide limited support for their answers.
Step 1	Students select pictured representations of ideas and identify stated details contained in simple texts. In written responses, they can select and transfer information from charts.

[a] Taken from CTB/McGraw-Hill (1997c), p. 51.

Of the three publishers, Riverside was the only one to provide specific content ALDs to the panelists at the beginning of the standard-setting process, as indicated by the statement below:

> Before workshop participants could be given the task of identifying performance levels [cut-scores] for the items of [the *Iowa Tests of Basic Skills*], they had to be given clear guidelines as to what should constitute different levels of performance on the tests . . . . The guidelines given to the workshop judges provided a brief description of the test to be rated. General performance descriptors followed, with the supplementary content statements bulleted under each general descriptor. Collectively, these statements would be used by the judges to place the specific content of each test item into the context of the general performance descriptors. (Riverside Publishing, 1998, p. 3)

Table 4 shows the descriptors that Riverside used for the grade 4 reading test. Presumably, the ALDs provided to the panelists by Riverside were fixed; that is, the descriptors could not be changed. The task for the standard-setting panelists was to translate these descriptors into cut-scores on the *Iowa Tests of Basic Skills* (*ITBS)* score scale.

Table 4. Riverside Sample Performance Descriptor Guidelines for Grade 4 Reading [a]

The *ITBS* Reading Comprehension test consists of a variety of reading materials. These include fiction, poetry, and nonfiction in the areas of social studies, science, and autobiography. All the poems and many of the other selections are excerpts from published literature. In interpreting the descriptions below, it should be kept in mind that the reading materials are assumed to be appropriate for grade 4 and that the descriptions cumulate across levels: Basic to Proficient to Advanced.

Advanced Fourth-grade students performing at the advanced level generalize about ideas and information in text that they read, and evaluate texts critically.

Students performing at the advanced level:
- extend the meaning of a text to other situations
- identify the author's point of view/purpose
- recognize aspects of mood/tone/style/structure
- identify meaning/purpose of nonliteral language

Proficient Fourth-grade students performing at the proficient level identify ideas and information suggested by, but not explicitly stated in, text they read.

Students performing at the proficient level:
- recognize cause-and-effect relationships
- predict likely outcomes
- draw appropriate conclusions
- determine the main ideas in a text

Basic Fourth-grade students performing at the basic level understand the overall literal meaning of text they read.

Students performing at the basic level:
- identify factual details
- understand literal meanings of words or phrases
- make simple inferences
- summarize a text

Below Basic Fourth-grade students performing at the below basic level do not meet the grade-level standard for basic achievement.

[a] Taken from Riverside Publishing (1998), p. 3.

Methods Used To Derive Cut-Scores

In this subsection, the specific procedures used by the three publishers to set the cut-scores associated with their achievement or performance levels are described. Information in this subsection is taken primarily from the following three sources:

1. *Technical Bulletin 1* (CTB/McGraw-Hill, 1997b).

2. *Content and Performance Standards* (Harcourt Brace, 1997a).

3. *Special Report on Riverside's National Performance Standards* (Riverside Publishing, 1998).

As noted in the preceding subsection, each of the three publishers developed some type of ALDs for use by the standard-setting panelists. All publishers also required panelists to take the assessments for which the cut-scores would be established. Each publisher began the standard-setting workshop with a general orientation session and used in-house content/technical personnel as facilitators. Throughout these standard-setting workshops, publishers provided panelists with training for each standard-setting activity. An overview of each publisher's standard-setting workshop is given in table 5. The remainder of this subsection will elaborate on the topics identified in table 5. Before beginning this elaboration, however, it may be useful to note that the standard-setting procedures used by Harcourt Brace and Riverside were somewhat similar both to one another and to the NAEP standard-setting procedures.[6]

[6] See footnote 3.

Table 5. Overview of Publishers' Standard-Setting Workshops

	CTB/McGraw-Hill	Harcourt Brace	Riverside
Date	Summer 1996	Summer 1995	Summer 1996
Length of Time	Over a 3-week period	1 week	3, 4, or 5 days
Number of Panelists From Outside Company	More than 50	More than 200	159
Geographic Distribution of Panelists	School districts from across the country	School districts from across the country	School districts from across the country
Occupations of Panelists	Teachers and curriculum experts	Teachers	Teachers, curriculum specialists, and department leaders
Grades	Four grade groups • Grades 1 and 2 • Grades 3, 4, and 5 • Grades 6, 7, and 8 • Grades 9, 10, 11, and 12	Grades 1 through 12	Grades 2, 4, 6, 8, 10, and 12
Major Content Areas	Reading, language, mathematics, social science, science	Reading, language, mathematics, social studies, science	Reading, language, mathematics, social studies, science
Items	Multiple choice and constructed response	Multiple choice and constructed response	Multiple choice and constructed response
Number of Panelists per Standard-Setting Panel	5 or 6	12 or 13	Minimum of 10
Method for Setting Cut-Scores	Bookmark	Modified Angoff	Modified Angoff
Number of Rounds	3	2	2
Normative Data	Items ordered by relative difficulty	Item *p*-values	Item *p*-values
Consequences Data	Yes	No	No

Length of Time

Given the information in the publications noted above, the exact length of time taken by a panel of judges to set cut-scores for a given grade and a given content area (e.g., grade 4 reading) could not be ascertained. It seems clear that individual judges usually served on more than one panel. A fourth-grade teacher, for example, might have served on a reading panel, a mathematics panel, and another panel.

Number of Panelists

In addition to the panelists from outside the company, all CTB/McGraw-Hill panels included a "CTB content expert" (CTB/McGraw-Hill, 1997b, p. 235). In-house content experts did not appear to be members of the standard-setting panels of the other two publishers. As noted previously, all publishers used in-house personnel as leaders and facilitators.

Geographic Distributions of Panelists

Although educators from throughout the country were part of all standard-setting panels, none of the publishers claims that its panels were representative of U.S. educators.

Occupations of Panelists

The publishers' standard-setting panelists consisted of teachers and curriculum specialists. Thus, the makeup of these panels differs from those convened for NAEP standard settings. NAEP panels also include members of the public.

Grades

The standard-setting panels convened by Harcourt Brace and Riverside set cut-scores within a given grade (e.g., grade 4). However, the CTB/McGraw-Hill panels set cut-scores across two or more grades (e.g., grades 3, 4, and 5). CTB/McGraw-Hill explains its decision to use multiple grades as follows:

> The five performance levels in each grade group are intended to be viewed as steps along a path toward the goal of proficient or advanced performance level placement by the time a student completes the highest grade in the grade group. That is, students who attain proficient or advanced placement in a particular grade group are considered to have attained the goals commonly set forth in the curriculum of the highest grade in that grade group.
>
> When using the performance levels as an indicator of student progress, both the performance level and the grade the student is in must be considered. For example, in the Elementary Grade Group [Grades 3, 4, and 5], the performance levels represent steps on the path to fifth-grade proficiency. Thus, the goal is to achieve proficient or advanced placement by the end of the fifth grade. One reason it was decided to set performance levels in across-grade groupings was to enable students (and their teachers and parents) to see their academic growth over time.

> If separate performance levels were set for each test level grade, chances are that most children would stay in the same level year after year. Thus, although students would grow from year to year, this growth would not be directly reflected by a commensurate movement to a higher performance level. By setting standards across grade levels, student growth will more likely be reflected by movement into higher performance levels as they move from one grade to the next. (CTB/McGraw-Hill, 1997b, pp. 234–235)

Obviously, whether the achievement level descriptions and the associated cut-scores are established within grade (Harcourt Brace and Riverside) or across grades (CTB/McGraw-Hill) has important implications when interpreting achievement level results. For example, using the CTB/McGraw-Hill procedure probably guarantees that, for a given grade group, a larger percentage of students in an upper grade will be in the highest achievement level relative to the percentage of students in a lower grade. Likewise, a smaller percentage of students in an upper grade will be in the lowest achievement level relative to the percentage of students in a lower grade. However, for the Harcourt Brace and Riverside procedures, no such systematic differences across a given set of grades would be predicted.

Major Content Areas

CTB/McGraw-Hill and Riverside only set cut-scores in the five areas noted in table 5. Harcourt Brace panelists also set cut-scores in several other areas (e.g., study skills and listening).

Items

All three publishers required panelists to consider both multiple-choice and constructed-response items as part of the standard setting. Although not specifically stated in *Technical Bulletin 1* (CTB/McGraw-Hill, 1997b), the CTB/McGraw-Hill panelists are assumed to have worked with items from tests targeted at all the grades in their grade group. Thus, for example, educators in the grades 3, 4, and 5 panel are assumed to have set the cut-scores using all the items from the tests targeted at grades 3, 4, and 5 (*TerraNova*, Levels 13, 14, and 15). The Harcourt Brace and Riverside panelists would have set the cut-scores using only items from a single test level of either the *Stanford* or the *ITBS*.

Method for Setting Cut-Scores

As noted in table 5, both Harcourt Brace and Riverside used what they labeled a modified Angoff procedure.[7] For multiple-choice items, both publishers asked panelists to estimate the percentage of borderline students (e.g., borderline Basic) who would be able to answer the item correctly. The directions they gave to panelists for constructed-response items differed, however. Harcourt Brace instructed its panelists "to assign percentages of each borderline group that should receive each of the score points . . ." (Harcourt Brace, 1997a, pp. 3 and 4). For each constructed-response (CR) item worth more than 1 point, Riverside instructed its panelists to identify "the typical score (or

[7] Labeling a procedure as "modified Angoff" only provides an indication of the general characteristics of the method used to set the cut-scores. There seems to be no single modified Angoff method.

most common number of points) earned by individuals" in the borderline group (Riverside Publishing, 1998, p. 5).[8]

For both Harcourt Brace and Riverside, the raw score cut-points for each panelist were obtained by summing the panelist's ratings across items within a test. These individual cut-scores were averaged across panelists to arrive at the group cut-score. This group cut-score on the raw score scale was then transformed to the primary score scale used by the publisher to report test results.

CTB/McGraw-Hill's method for setting cut-scores, the bookmark standard-setting procedure, differs markedly from the modified Angoff procedure. The first step in the bookmark procedure is to order the test items along the primary score scale used with the test. The location of a multiple-choice item on this score scale "is defined for the purposes of standard setting as the point on the score scale at which a student would have a two-thirds (0.67) probability of success (with guessing factored out)" (CTB/McGraw-Hill, 1997b, p. 236).[9] For CR items with more than two possible score values, "[T]he locations of each CR item score point is defined as the position on the score scale at which students with the given ability level have a two-thirds probability of achieving at least that score point, that is, that score point or better" (CTB/McGraw-Hill, 1997b, p. 236).

Once the items (or CR score points) have been ordered, an "ordered-item booklet" is prepared. Each item (or score point) appears on a separate page of this booklet. Panelists then place a "bookmark" in the booklet to identify the "unique cut-score for a given performance level" (CTB/McGraw-Hill, 1997b, p. 236). Items preceding the bookmark represent content that all students at that performance level should know. The cut-score for a particular performance level is "computed as the mean of the location (on the score scale) of the items immediately before and after the bookmark" (CTB/McGraw-Hill, 1997b, p. 236).

Number of Rounds

All three publishers required panelists to work independently during the first round of item rating or bookmarking. At the end of this round, the individual panelist's cut-scores (bookmarks) were provided to the panels. The Harcourt Brace and Riverside panels also received the group cut-scores for their consideration. Then, a second round of independent ratings or bookmarks was obtained from panelists. For the Harcourt Brace and Riverside panels, these second-round item ratings were used to establish the final cut-scores. For the CTB/McGraw-Hill panelists, this second set of bookmarks was used to provide panelists with individual and group cut-scores on the score scale. This information and some "consequences data" were given to panelists before a third and final round of bookmarking. The final cut-scores were based on the panelists' bookmarks for round 3.

---

[8] Both of these procedures differ somewhat from the procedures used with recent NAEP assessments. In the 1998 NAEP Civics Assessment, panelists were asked to estimate the mean score for the borderline group on the CR items (Loomis et al., 1999).

[9] The specific procedures used to "factor out guessing" are not provided in *Technical Bulletin 1* (CTB/McGraw-Hill, 1997b).

Normative Data

At the end of round 1, the Harcourt Brace and Riverside panelists were given item-difficulty information for the items they were rating. The CTB/McGraw-Hill Bookmark procedure actually requires normative data to prepare the "ordered-item booklets." The location of the items in the booklet indicates the relative level of difficulty for the items. More specifically, the location provides information about the probability that examinees at that ability level will answer the multiple-choice item correctly. As noted above, CTB/McGraw-Hill decided to locate the items at a point on the proficiency scale where this probability was approximately .67 ("with guessing factored out").[10] Thus, for a given multiple-choice item, Harcourt Brace and Riverside provided panelists with an estimate of the proportion (*p*-value) of a national grade-level sample that would answer the item correctly and CTB/McGraw-Hill provided panelists with an estimate of the proportion (2/3) of a subgroup of the national sample that would answer the item correctly.[11]

Consequences Data

Of the three publishers, only CTB/McGraw-Hill provided consequences (or impact) data to panelists. Before round 3 bookmarks, its panelists were provided with "the percentage of students nationally [who] would be expected to fall in each performance level based on the group bookmark placements [after round 2]" (CTB/McGraw-Hill, 1997b, p. 238). This information was considered by the panelists before they made their round 3 bookmarks.

Outcomes of the Standard-Setting Process

For many standard-setting projects, the two major outcomes of interest are Final Achievement Level Descriptions (FALDs) and cut-scores. In some situations, a third outcome, the identification of exemplary items from the item pool to help interpret the FALDs, is also of interest. However, such exemplary items cannot be provided with the FALDs developed by the three publishers because the items in the item pool are included in tests currently being administered in schools.

The cut-scores from a standard setting are used by the publishers primarily to develop standards-based reports for individuals and groups. Interpretation of such reports is enhanced if the publisher also provides information about the percentages of a representative national sample at the achievement levels. All three publishers provide this type of information.

In this subsection, the procedures used by the publishers to arrive at their FALDs are considered briefly. Then, for one content area (reading comprehension) and two grades (4 and 8), the estimates of the national percentages of students at the achievement levels are reported for each publisher. A brief discussion of these percentages concludes the subsection.

[10] It should be noted that the relative order of the items could change if a different probability value was selected.

[11] This subgroup is defined by a score on the *TerraNova* score scale and is not a within grade-level group.

Final Achievement Level Descriptions

As noted in the first subsection of this paper, Riverside provided ALDs to its panelists at the beginning of the standard-setting process. These descriptions presumably remained unchanged after the cut-scores were set. Thus, Riverside's FALDs were, in fact, its initial ALDs. An example of these descriptions is given in table 4.

Both CTB/McGraw-Hill and Harcourt Brace began their standard-setting process with initial ALDs developed by panelists. These initial descriptions were modified at the end of the process, after the cut-scores for the achievement levels were established.

An example of the FALDs reported by CTB/McGraw-Hill is shown in table 3. A brief summary of how these FALDs were developed is given below:

> After the final cut scores were determined, the [panel's] initial performance level descriptors had to be modified to their final form. This was rather straightforward; the participants' final cut scores were translated into corresponding final bookmark locations. Items prior to a given final bookmark in the ordered-item booklet (items with scale locations at or below the corresponding final cut score) represent what all students in the corresponding performance level are expected to know and perform. The content of these items was synthesized to generate the descriptors for each performance level.[12] (CTB/McGraw-Hill, 1997b, p. 239)

As can be seen in table 3, the FALDs for reading in grades 3 through 5 contain statements related to performance on both multiple-choice items and constructed-response items.

As noted previously, Harcourt Brace provided its standard-setting panelists with policy definitions of four achievement levels. The panelists then derived operational definitions of the achievement levels for each grade and content area based on these policy definitions and the test content.[13] Although operational definitions of the ALDs were developed, it appears that only the policy definitions are part of Harcourt Brace's standards-based reports. Harcourt Brace, however, does provide a more detailed description of what students at a given achievement level can do. These additional descriptions were developed using the following procedures:

> Once the cut points for the Performance Standards [achievement levels] had been determined, we reanalyzed the standardization data in order to assign the students in the standardization to the appropriate Performance Standard category and then look at the results for each of the groups. For each question in each subtest, we calculated *p*-values—the percentage of students answering the question correctly—for each of the Performance Standard groups, that is, the percentage of students in Level 4 answering the question correctly, the percentage of students in Level 3 answering it correctly, and the percentage of students in

[12] These final descriptions can be found in *TerraNova Performance Levels Handbook* (CTB/McGraw-Hill, 1997c).

[13] As noted previously, these operational definitions could not be located.

> Level 2 answering it correctly. Thus we can examine the difficulty of the test questions for students whose performance puts them into the Advanced category, for students who fall into the Proficient category, and for students whose performance is in the Basic category, i.e., partial mastery. And since we know the kinds of things the test questions are asking students to do, we can make generalizations about the things students at each of the Performance Standard categories are able to do, based on their ability to answer the test questions.
>
> In order to condense the data, we reasoned that for any given test question, a *p*-value of 0.80 (or above) would represent a group's "ability to do what the question is asking." Therefore, the data in this report are based on percentages of test questions that have *p*-values of 0.80 or greater. In other words, the concept we're reporting here is the percentage of test questions, or content, students at each Performance Standard level can do. (Harcourt Brace, 1997b, p. 2)

Table 6 shows the outcomes of this process for one of the *Stanford* reading comprehension tests. To illustrate how the information in table 6 can be interpreted, the data for the "Critical Analysis" row will be considered. The reading comprehension test for this level of the *Stanford* contains nine items that have been classified as critical analysis items (Harcourt Brace, 1996b, p. 79). From table 6, it can be seen that the percentage of these items that have been mastered by students at the Advanced level is 100. That is, for the group of students classified as Advanced, the *p*-values for these nine items are all greater than or equal to 0.80. For students at the Proficient level, 67% of the nine items (six items) have *p*-values of 0.80 or greater and for students at the Basic level, 22% of the nine items (two items) have *p*-values of this magnitude.[14]

The procedures used by CTB/McGraw Hill and Harcourt Brace to describe what students at a particular achievement level can do seem to be similar to the scale anchoring procedures used with NAEP assessments. (See, for example, Campbell et al., 1996.)

Achievement Level Percentages: Reading, Grades 4 and 8

Tables 7 and 8 show the percentages of students (grades 4 and 8) in the publishers' national samples who were classified at each reading achievement level. These tables also report the estimated national percentages of students at the four NAEP achievement levels.[15] These NAEP percentages are based on the results of the 1994 NAEP Reading Assessment.

[14] The percentages reported in table 6 are for the multiple-choice items in this test level.

[15] Tables 9 through 12 in appendix A provide similar data for mathematics and science.

Table 6. Harcourt Brace: Intermediate 1 Reading Comprehension[a, b]

	Percentage of Content Mastered at Each Performance Level			
	Level 1 Below Basic	Level 2 Basic	Level 3 Proficient	Level 4 Advanced
The **Reading Comprehension** subtest is composed of reading selections accompanied by questions about each selection.	<27	27	64	90
Recreational: Read literature for enjoyment of literary merit, including folk tales, historical fiction, contemporary fiction, humor, and poetry.	<39	39	89	100
Textual: Read level-appropriate expository material, with content from the natural, physical, and social sciences, as well as other nonfiction general information materials.	<22	22	56	83
Functional: Read material encountered in everyday life, both inside and outside of school, including directions, forms, schedules, personal notes, and advertisements.	<22	22	50	83
Within each type of text, questions measure reading achievement in four modes of comprehension				
Initial understanding: Understand explicitly stated information—grasp details, actions, behaviors, sequences of events.	<42	42	75	100
Interpretation: Make inferences from explicit and implicit information in the text, and make generalizations from this information. This can include interpreting the author's use of figurative language; interpreting the main idea of a selection; determining cause and effect relationships; and making predictions.	<29	29	67	83
Critical Analysis: Synthesize and evaluate explicit and implicit information in a variety of reading selections.	<22	22	67	100
Strategies: Recognize and apply text factors and reading strategies in a variety of reading selections.	<11	11	44	78

[a] The recommended grade range for Intermediate 1 is 4.5–5.9.

[b] Adapted from Harcourt Brace (1997b), p. 10.

Table 7. Estimated National Percentages of Students at Each Achievement Level: Reading Comprehension Grade 4 Spring

CTB/McGraw-Hill[a]		Harcourt Brace[b]		Riverside[c]		NAEP[d]	
Level	%	Level	%	Level	%	Level	%
Advanced	7	Advanced	13	Advanced	17	Advanced	7
Proficient	17	Proficient	27	Proficient	38	Proficient	23
Nearing Proficiency	32	Basic	30	Basic	31	Basic	30
Progressing	24	Below Basic	30	Below Basic	14	Below Basic	40
Step 1	20						

[a] From CTB/McGraw-Hill, 1997b, p. 243.
[b] From Harcourt Brace, 1997c, p. 461.
[c] From Riverside Publishing, 1998, p. 9.
[d] From Campbell et al., 1996, p. 44.

Table 8. Estimated National Percentages of Students at Each Achievement Level: Reading Comprehension Grade 8 Spring

CTB/McGraw-Hill[a]		Harcourt Brace[b]		Riverside[c]		NAEP[d]	
Level	%	Level	%	Level	%	Level	%
Advanced	10	Advanced	6	Advanced	15	Advanced	3
Proficient	23	Proficient	33	Proficient	32	Proficient	27
Nearing Proficiency	29	Basic	34	Basic	35	Basic	40
Progressing	19	Below Basic	27	Below Basic	18	Below Basic	30
Step 1	19						

[a] From CTB/McGraw-Hill, 1997b, p. 243.
[b] From Harcourt Brace, 1997c, p. 465.
[c] From Riverside Publishing, 1998, p. 9.
[d] From Campbell et al., 1996, p. 44.

A few observations and comments related to the interpretation of the data in tables 7 and 8 are noted below:

1. The percentages reported for CTB/McGraw-Hill, Riverside, and NAEP are based on content domains that include both constructed-response items and multiple-choice items. The percentages given for Harcourt Brace are for domains consisting only of multiple-choice items. Harcourt Brace also reported these percentages for constructed-response items only.[16] For grade 4, these percentages were: 4% (Advanced), 23% (Proficient), 41% (Basic), and 32% (Below Basic). For grade 8, the percentages were: 3% (Advanced), 26% (Proficient), 41% (Basic), and 30% (Below Basic) (Harcourt Brace, 1997c, pp. 461 and 465).[17]

2. The CTB/McGraw-Hill standard-setting panelists set cut-scores for grade groups (e.g., grades 3, 4, and 5), whereas the Harcourt Brace, Riverside, and NAEP panelists set standards for a single grade (e.g., grade 4). For CTB/McGraw-Hill, grade 4 was the middle grade of three grades and grade 8 was the highest grade of three grades. As noted earlier in this section, the use of multiple-grade panels creates a predictable relationship between the grade level within a panel's group and the percentage of students at the highest and lowest achievement levels. Thus, for example, because grade 8 is the highest grade in a three-grade group, it is not surprising that the percentage of students at the Advanced level in grade 8 (10%) is greater than the percentages for grade 7 (7%) and grade 6 (4%) (CTB/McGraw-Hill, 1997b, p. 43).[18] Because the Harcourt Brace and Riverside panels were setting cut-scores for a single grade, such a relationship would not necessarily be observed (Harcourt Brace, 1997c, pp. 459–469; Riverside Publishing, 1998, p. 9).

3. The percentages reported in tables 7 and 8 are based on national samples from 3 different years. The CTB/McGraw-Hill data are for 1996, the Harcourt Brace and Riverside data are for 1995, and the NAEP data are for 1994.[19]

---

[16] These are labeled open-ended items by Harcourt Brace.

[17] A comparison of these percentages and the percentages in tables 7 and 8 indicates that the cut-scores for the constructed-response items are higher than the cut-scores for the multiple-choice items. This outcome is consistent with the outcomes of the 1992 NAEP Reading Assessment (National Academy of Education, 1996, p. 94).

[18] This procedure also leads to predictable changes in the percentages across grade groups. To illustrate the nature of these changes, consider the data shown below:

	Grade Group 1			Grade Group 2		
	3	4	5	6	7	8
Percentage at Advanced	3	7	9	4	7	10
Percentage at Two Lowest Levels	58	44	30	55	46	38

As these data illustrate, the lowest grade in the upper grade group (grade 6) has a smaller percentage of students at the Advanced level and a larger percentage of students at the two lowest levels relative to these percentages in the highest grade in the lower grade group (grade 5).

[19] The percentages for the 1998 NAEP Reading Assessment are very similar (Donahue et al., 1999).

4. The NAEP percentages are derived from estimated true score distributions, whereas the publishers' percentages are based on observed score distributions. If the publishers had derived these percentages for estimated true score distributions, the values reported for both the highest and lowest levels would have decreased. However, given the relatively high reliability of these tests, the decrease would be small.[20]

5. As can be seen in tables 7 and 8, the percentage of students at a given achievement level (e.g., Advanced) varies substantially across publishers. Because three of the four sets of data in tables 7 and 8 have the same labels for all achievement levels, it is tempting to speculate about the reasons for the observed differences among percentages. However, it would be extremely difficult, if not impossible, to draw any firm conclusions either about why such differences occur or about the implications of such differences.[21]

   In addition to the differences among the publishers' standard-setting procedures noted in the four comments above, other differences are easily identified. Their test specifications differ; thus, their content domains are not the same.[22] Compare, for example, the descriptions of the Advanced achievement levels given in tables 3 and 4. Their methods used to establish cut-scores differ. (These differences were considered in some detail in the previous subsection.) Finally, the use of policy definitions and achievement level descriptions also differs across publishers.[23, 24] Given such differences, perhaps the variation among the percentages reported in tables 7 and 8 should not be considered surprising.

---

[20] Consider, for example, the Harcourt Brace and Riverside percentages reported in table 7 for the Advanced level and the Below Basic level. Given the reliabilities of these tests and assuming a normal distribution for the estimated true scores, these percentages would have decreased by approximately 1% if estimated true score distributions had been used instead of observed score distributions.

[21] A similar caution is noted by Riverside Publishing (1998, p. 9): "The National Assessment of Educational Progress (NAEP) uses similar labels for its performance categories. However, because the two sets of performance standards are based on different tests and different standard-setting research, they cannot be considered interchangeable."

[22] It should also be noted that the CTB/McGraw-Hill content domains used for standard setting cover more than one grade.

[23] Neither CTB/McGraw-Hill nor Riverside used policy definitions. Harcourt Brace used policy definitions that were somewhat similar to the NAEP definitions, although not identical. For example, the Harcourt Brace policy definition for Proficient includes the qualifying phrase "indicating that students are prepared for the next grade." The NAEP definition for Proficient does not contain this phrase.

[24] CTB/McGraw-Hill used five achievement levels, whereas the other publishers used four.

Concluding Statement

In this section, the standard-setting procedures used by the publishers of three standardized achievement tests were described briefly. These publishers have established standards in several content areas; however, this section focused only on reading. At times, the standard-setting procedures used with some NAEP assessments were also discussed. This section did not evaluate the relative merits of the different standard-setting procedures.

Although the various procedures had some similar characteristics, the differences among the procedures were substantial. Thus, it was not surprising that despite the use of similar labels for the achievement levels, the publishers' final achievement level descriptions and the percentages of students at the various levels differed considerably.

References

Bourque, M.L., Champagne, A.B., and Crissman, S. (1997). *1996 Science performance standards: Achievement results for the Nation and the States*. Washington, DC: National Assessment Governing Board.

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

Campbell, J.R., Donahue, P.L., Reese, C.M., and Phillips, G.W. (1996). *NAEP 1994 reading report card for the Nation and the States*. Washington, DC: National Center for Education Statistics.

CTB/McGraw-Hill (1997a). *TerraNova*. Monterey, CA.

CTB/McGraw-Hill (1997b). *Technical Bulletin 1*. Monterey, CA.

CTB/McGraw-Hill (1997c). *TerraNova Performance Levels Handbook*. Monterey, CA.

CTB/McGraw-Hill (1999). *Teacher's Guide to TerraNova*. Monterey, CA.

Donahue, P.L., Voelkl, K.E., Campbell, J.R., and Mazzeo, J. (1999). *NAEP 1998 reading report card for the Nation and the States*. Washington, DC: National Center for Education Statistics.

Hanick, P.L., and Loomis, L.C. (1999). *Setting standards for the 1998 NAEP in civics and writing: Using focus groups to finalize the achievement level descriptions.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal.

Harcourt Brace (1996a). *Stanford Achievement Series: Ninth edition.* San Antonio, TX.

Harcourt Brace (1996b). *Stanford Achievement Series, ninth edition: Compendium of instructional objectives*. San Antonio, TX.

Harcourt Brace (1997a). *Content and performance standards.* San Antonio, TX.

Harcourt Brace (1997b). *Performance standard scores.* San Antonio, TX.

Harcourt Brace (1997c). *Stanford Achievement Series, ninth edition: Technical data report.* San Antonio, TX.

Hoover, H.D., Hieronymus, A.N., Frisbie, D.A., and Dunbar, S.B. (1996a). *Iowa tests of basic skills, Form M*. Itasca, IL: Riverside Publishing.

Hoover, H.D., Hieronymus, A.N., Frisbie, D.A., and Dunbar, S.B. (1996b*). Iowa tests of basic skills: Interpretive guide for teachers and counselors, Form M.* Itasca, IL: Riverside Publishing.

Loomis, S.C., Bay, L., Yang, W., and Hanick, P.L. (1999). *Field trials to determine which rating method to use in the 1998 NAEP achievement levels-setting process for civics and writing.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal.

National Academy of Education (1996). *Quality and utility: The 1994 trial State assessment in reading*. Stanford, CA.

National Assessment Governing Board (1998). *Civics assessment framework for the 1998 National Assessment of Educational Progress*. Washington, DC.

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.

Riverside Publishing (1998). *Special report on Riverside's national performance standards.* Itasca, IL.

Appendix A

Table A1. Estimated National Percentages of Students at Each Achievement Level: Mathematics Grade 4 Spring[a]

CTB/McGraw-Hill[b]		Harcourt Brace[c]		Riverside[d]		NAEP[e]	
Level	%	Level	%	Level	%	Level	%
Advanced	3	Advanced	7	Advanced	14	Advanced	2
Proficient	12	Proficient	27	Proficient	30	Proficient	19
Nearing Proficiency	32	Basic	39	Basic	39	Basic	43
Progressing	30	Below Basic	27	Below Basic	17	Below Basic	36
Step 1	23						

[a] For Harcourt Brace and Riverside, the percentages are for Mathematics Total.
[b] From CTB/McGraw-Hill, 1997b, p. 243.
[c] From Harcourt Brace, 1997c, p. 461.
[d] From Riverside Publishing, 1998, p. 9.
[e] From Reese, et al., 1997, p. 47.

Table A2. Estimated National Percentages of Students at Each Achievement Level: Mathematics Grade 8 Spring[a]

CTB/McGraw-Hill[b]		Harcourt Brace[c]		Riverside[d]		NAEP[e]	
Level	%	Level	%	Level	%	Level	%
Advanced	10	Advanced	3	Advanced	8	Advanced	4
Proficient	23	Proficient	20	Proficient	29	Proficient	20
Nearing Proficiency	29	Basic	35	Basic	42	Basic	38
Progressing	20	Below Basic	42	Below Basic	21	Below Basic	38
Step 1	18						

[a] For Harcourt Brace and Riverside, the percentages are for Mathematics Total.
[b] From CTB/McGraw-Hill, 1997b, p. 243.
[c] From Harcourt Brace, 1997c, p. 465.
[d] From Riverside Publishing, 1998, p. 9.
[e] From Reese, et al., 1997, p. 47.

Table A3. Estimated National Percentages of Students at Each Achievement Level: Science Grade 4 Spring

CTB/McGraw-Hill[a]		Harcourt Brace[b]		Riverside[c]		NAEP[d]	
Level	%	Level	%	Level	%	Level	%
Advanced	5	Advanced	10	Advanced	16	Advanced	3
Proficient	19	Proficient	32	Proficient	39	Proficient	26
Nearing Proficiency	33	Basic	38	Basic	28	Basic	38
Progressing	27	Below Basic	20	Below Basic	17	Below Basic	33
Step 1	16						

[a] From CTB/McGraw-Hill, 1997b, p. 243.
[b] From Harcourt Brace, 1997c, p. 461.
[c] From Riverside Publishing, 1998, p. 9.
[d] From Bourque, et al., 1997, p. viii.

Table A4. Estimated National Percentages of Students at Each Achievement Level: Science Grade 8 Spring

CTB/McGraw-Hill[a]		Harcourt Brace[b]		Riverside[c]		NAEP[d]	
Level	%	Level	%	Level	%	Level	%
Advanced	9	Advanced	7	Advanced	16	Advanced	3
Proficient	25	Proficient	21	Proficient	33	Proficient	26
Nearing Proficiency	25	Basic	35	Basic	33	Basic	32
Progressing	24	Below Basic	37	Below Basic	18	Below Basic	39
Step 1	17						

[a] From CTB/McGraw-Hill, 1997b, p. 243.
[b] From Harcourt Brace, 1997c, p. 465.
[c] From Riverside Publishing, 1998, p. 9.
[d] From Bourque, et al., 1997, p. viii.

SECTION 5

States With NAEP-Like Performance Standards

Jeffrey M. Nellhaus Massachusetts Department of Education

November 2000

States With NAEP-Like Performance Standards

Jeffrey M. Nellhaus

In recent years, there has been a profusion of Federal and State mandates aimed at improving student achievement in the Nation's public elementary and secondary schools. In response to these new requirements, many States developed academic standards in core content areas and student assessment programs, based on those academic standards, that are designed to measure the extent to which students are achieving them. Consistent with the purpose of these assessments, States also developed achievement level categories, like those used by the National Assessment of Educational Progress (NAEP), to report results. In many States, committees of educators and other stakeholder groups were convened to define performance expectations at each achievement level. In turn, the definitions were used to determine the scores associated with each of the achievement levels through a process called standard setting. The names, definitions, and test scores concomitant with these achievement levels are commonly referred to as performance standards.

The purpose of this section is to: (1) summarize recent State-level efforts to develop performance standards, and (2) determine the similarity of States' performance standards to those adopted by the National Assessment Governing Board (NAGB) for reporting the results of NAEP. This information is intended to assist NAGB in reviewing the performance standards currently used to report NAEP results and to inform discussions of how those performance standards might be refined in the future. Specifically, this section attempts to answer the following questions:

1. In which States do State assessment directors perceive the achievement level categories used to report the results of their assessments to be the same as or similar to NAEP?

2. In which States is the number of achievement level categories the same as or similar to NAEP?

3. In which States are the names of achievement level categories the same as or similar to NAEP?

4. In which States are the descriptions of achievement level categories the same as or similar to NAEP?

5. In which States was the method used to set standards the same as or similar to NAEP?

6. In which States is the combination of the number of achievement level categories, the names of categories, the descriptions of categories, and the standard-setting method similar enough to NAEP that the State's performance standards can be characterized as NAEP-like?

7. The last portion of this section explores whether States report assessment results consistent with State NAEP. The assumption is that if a State assessment program and State NAEP are similar (in terms of what they measure, how they measure it, and how they report results), then the results reported by the two assessment programs should be similar. The findings

presented in this portion of the section are based on a small sample of State data and should be interpreted with caution.

Sources of Information, Method, and Findings

Sources of Information

Information in this report was derived from several sources. Information about State assessment programs was obtained from *The Annual Survey of State Student Assessment Programs, Fall 1998* (Council of Chief State School Officers 1998). Data not available for some States in the Council of Chief State School Offices (CCSSO) *Annual Survey* were obtained by telephone conversations and e-mail communications with State assessment directors and from a wide array of reports posted on the World Wide Web by State educational agencies. In addition, a questionnaire was sent by e-mail to all State assessment directors. State-level NAEP results were obtained from the *NAEP 1998 Reading Report Card* (National Center for Education Statistics, 1999) and the *NAEP 1996 Mathematics Report Card* (National Center for Education Statistics, 1997).

Method

Two methods were used to determine whether a State has NAEP-like performance standards. First, State assessment directors were requested to complete a questionnaire to determine whether they perceived their State's performance standards to be similar to NAEP. Second, the component parts of each State's performance standards were identified and subsequently classified as being the same as, similar to, or different from the corresponding NAEP component. The components of each State's performance standards were then considered collectively to determine whether they are similar to NAEP overall.

State Assessment Director Questionnaire. A brief questionnaire was sent via e-mail to all State testing directors in August 1999 asking: "Are the results of any of your State's assessments reported using achievement level categories that are similar to or the same as those used by NAEP? Please answer for assessments you administered during the 1998–1999 school year or plan to administer during the 1999–2000 school year." Respondents could answer "yes," "no," or "not sure." The questionnaire was e-mailed to 50 State testing directors and the testing directors in Puerto Rico and the U.S. Virgin Islands.

Results of Questionnaire. Of the 34 State directors who responded to the questionnaire, 17 (50%) responded yes, 12 (35%) responded no, and 5 (15%) responded not sure. Eighteen directors did not respond. Responses to the survey by State are summarized in table 1.

Table 1. Results of State Assessment Director Questionnaire

Responses of State Assessment Directors to the Question Asking Whether Their State's Achievement Levels Are NAEP-Like			
Similar to NAEP (17)	**Not Similar to NAEP (12)**	**Not Sure (5)**	**No Response (18)**
Alaska	Alabama	Illinois	Arizona
Colorado	California	Kansas	Arkansas
Delaware	Connecticut	North Carolina	Georgia
Kentucky	Florida	Rhode Island	Idaho
Louisiana	Hawaii	Wyoming	Indiana
Maine	New York		Iowa
Massachusetts	North Dakota		Maryland
Michigan	Ohio		Mississippi
Minnesota	South Carolina		Nebraska
Missouri	Texas		Nevada
Montana	Virginia		New Hampshire
New Mexico	West Virginia		New Jersey
Oklahoma			Oregon
South Dakota			Pennsylvania
Utah			Puerto Rico
Washington			Tennessee
Wisconsin			U.S. Virgin Islands
			Vermont

Analysis of the Components of State Performance Standards. The second method used to determine whether a State's performance standards are similar to NAEP involved an analysis of the component parts of each State's performance standards. Accordingly, the following steps were carried out:

1. The components of each State's performance standards were identified, including the number of achievement level categories, the names of the categories, the descriptions of the categories, and the method used to set standards (to determine cut-scores).

2. For each State, each component was classified the same as, similar to, or different from NAEP.

3. The components were considered collectively and classified the same as, similar to, or different from NAEP.

Rules for Classifying the Components of State's Performance Standards. The following rules were applied to classify the various components of each State's performance standards:

Number of Achievement Levels

- If the State used four achievement level categories, the number of categories was classified the same as NAEP.
- If the State used more or fewer than four achievement level categories, the number of categories was classified different from NAEP.

Name of Achievement Levels

- If the names of the State's achievement level categories were identical to those of NAEP (Advanced, Proficient, Basic, and Below Basic), they were classified the same as NAEP.
- If the names of the State's achievement level categories were comparable to the names used by NAEP (e.g., Distinguished instead of Advanced, or Partially Proficient instead of Basic), they were classified similar to NAEP.
- If the names of the State's achievement level categories were expressed in terms of numbers or Roman numerals and used no other qualitative language (e.g., Level 1, Level 2, Level 3, and Level 4), they were classified different from NAEP.

Achievement Level Descriptions

- If the State's achievement level descriptions contained language comparable to the general policy definitions used by NAEP (Advanced means superior performance, Proficient means solid performance, Basic means partial understanding, and Below Basic means minimal understanding), they were classified similar to NAEP.
- If the State's achievement level descriptions were based on normative criteria or indicated distinctly higher or lower performance expectations than the corresponding NAEP achievement level descriptions, they were classified different from NAEP.

Standard-Setting Method

- If the State used the modified Angoff method, its standard-setting method was classified the same as NAEP, even though States may have implemented the method somewhat differently than NAEP.
- If the State used the bookmark or booklet classification methods or any other procedure that required individuals to make judgments about the difficulty of test items or the quality of student work in reference to the State's achievement level descriptions, the method was classified similar to NAEP.
- If the individuals setting the standards relied primarily on impact data (e.g., distribution of actual test score results), or if the cut-scores corresponded directly with a scoring rubric only (as was the case in a number of States' writing assessment programs), the standard-setting method was classified different from NAEP.

Overall Performance Standards

- A State's performance standards were classified similar to NAEP overall if they met the following criteria:
  - The achievement level definitions were classified similar to NAEP;
  - The method used to set cut-scores was classified the same as or similar to NAEP; and
  - Either the number or the names, or both the number and the names, of the achievement levels were classified the same as or similar to NAEP.

These criteria are summarized in table 2.

Table 2. Conditions Under Which State Performance Standards Were Classified To Be Similar to NAEP Performance Standards

Component of Performance Standards	All Components Same or Similar	All Components Same or Similar Except Level Number	All components Same or Similar Except Level Name
Names of Achievement Levels	✓	✓	Not the same or similar
Number of Achievement Levels	✓	Not the same or similar	✓
Description of Achievement Levels	✓	✓	✓
Standard-Setting Method	✓	✓	✓
Components Considered Collectively	✓	✓	✓

Results of Analysis of Component Parts of State Performance Standards. Appendix A provides for each State a description of the State assessment programs that designate achievement level categories for reporting results, the names of the categories, and the standard-setting method used to determine cut-scores. Appendix A also indicates whether each of the components of the State's performance standards are the same as, similar to, or different from NAEP, as determined by applying the aforementioned decision rules. The last column of the table indicates whether the State's performance standards are NAEP-like overall.

Tables 3 and 4 summarize the information presented in appendix A. Of the 52 jurisdictions (50 States, plus the U.S. Virgin Islands and Puerto Rico) included in this study, 23 were classified as having performance standards that are similar to NAEP, while 23 were classified as having standards that are different from NAEP. There was not enough information to classify the performance standards for six jurisdictions.

A summary of the standard-setting methods used by the States and territories is summarized in table 5. Most States used one of three methods: modified Angoff (6), bookmark (18), or booklet classification (8). Although the testing contractors employed by each State are not identified in this study, it is interesting to note that the standard-setting method used by each State appears to be related to the State's contractor. For example, States contracting with CTB McGraw Hill tend to use the bookmark method, States contracting with Harcourt Brace tend to use the modified

Angoff method, and States contracting with Advanced Systems tend to use a form of the booklet classification method referred to by Advanced Systems as the body of work method.

Table 3. Number of States With Performance Standard Components and Overall Performance Standards That Are Same as, Similar to, or Different From NAEP

Performance Standard Component	Similarity to NAEP			
	Same	Similar	Different	Not Sure
Number and Name of Levels	4*	33**	12***	3
Description of Levels	0	25	17	10
Standard-Setting Method	6	25	15	6
Overall	**0**	**23**	**23**	**6**

* Name and number are the same.
** Name and number are similar or either the name or number is similar.
*** Name and number are different.

Table 4. States With Performance Standards Similar to and Different From NAEP

Similar to NAEP (23)	Different from NAEP (23)	Insufficient Information (6)
Alaska	Alabama	California
Arizona	Connecticut	Nebraska
Arkansas	Florida	South Dakota
Colorado	Georgia	U.S. Virgin Islands
Delaware	Hawaii	Utah
Illinois	Idaho	West Virginia
Kentucky	Indiana	
Louisiana	Iowa	
Maine	Kansas	
Massachusetts	Maryland	
Michigan	Mississippi	
Minnesota	Montana	
Missouri	Nevada	
New Hampshire	New Jersey	
New Mexico	New York	
Oklahoma	North Carolina	
Pennsylvania	North Dakota	
Rhode Island	Ohio	
Tennessee	Oregon	
Vermont	Puerto Rico	
Washington	South Carolina (will be)	
Wisconsin	Texas	
Wyoming	Virginia	

Table 5. Standard-Setting Methods Used by States

Standard-Setting Method				
Modified Angoff (6)	Bookmark (18)	Booklet Classification (8)	Other (14)	Information Not Available or in Planning Phase (6)
California Connecticut Illinois Minnesota South Dakota Virginia	Alaska (may use) Arizona Colorado Delaware Indiana Louisiana Michigan* Missouri New Mexico New York Oklahoma Oregon Pennsylvania (may use) Tennessee Utah (may use) Vermont Washington Wisconsin	Arkansas Kentucky Maine Massachusetts New Hampshire New Jersey Rhode Island* Wyoming	<u>Scoring Guide</u> Alabama Florida Georgia Nevada <u>Contrasting Groups</u> North Carolina <u>Other</u> Hawaii Idaho Iowa Kansas Maryland Montana North Dakota Ohio Texas	Mississippi Nebraska Puerto Rico South Carolina U.S. Virgin Islands West Virginia

* Also use modified Angoff.

Comparison of State Assessment Results and State NAEP Results

Another test of whether States have developed standards that are similar to NAEP is to compare student performance on individual State assessments to their performance on comparable State NAEP assessments. The assumption is that if State assessments and comparable State NAEP assessments measure similar content using similar methods and use similar performance standards to report results, then the results of the two assessments should be similar.

Individual State assessment results and comparable State NAEP results are shown for nine States in tables 7 through 15, with a summary of the results of the comparisons shown in table 6. The tables depict results for a total of 15 assessments in the areas of reading and mathematics in grades 4 and 8. The various State assessments shown in tables 7 through 15 were selected because they were described earlier in this report to have performance standards that are similar to NAEP. The tables also indicate the standard-setting method used by the State.

To determine whether the results of the state and comparable State NAEP assessments shown in tables 7 through 15 are similar, first, the probable range in the percentage of students reported at each achievement level was determined by using the standard error associated with each assessment program. Specifically, two standard errors were used to recognize differences in the content, methods, and other factors associated with each assessment program. For example, if an assessment reported 10% at the Advanced level with a standard error of 1%, then the probable range in the percentage of students was determined to be 8–12%.

Subsequently, when the ranges in the percentage of students in corresponding achievement levels were compared and found to meet or overlap, the results on the two assessments at that level were characterized as similar. If they did not meet or overlap, the results were characterized as different.

Rounding to the nearest percentage point, NAEP typically reports a standard of error of 1 percentage point at the Advanced achievement level, and a standard error of 2 percentage points at the Proficient, Basic, and Below Basic levels. It is assumed here that the standard error for the State assessments is the same as NAEP. Accordingly, when the results of the assessments at the Advanced level were within four percentage points, they were characterized as similar. Results at the Proficient, Basic, and Below Basic levels were characterized as similar if they fell within 8 percentage points.

For the nine States shown in tables 7 through 15, a total of 60 comparisons were conducted (15 assessments, 4 achievement levels each). In more than half the cases (55%), State and State NAEP assessments reported similar results. In general, more instances of similarity were found at the Proficient and Basic levels than at the Advanced and Below Basic levels. The fewest instances of similarity occurred at the Below Basic level. State assessments tended to report a higher percentage of students performing at the Advanced level and a lower percentage of students performing at the Below Basic level than State NAEP.

Table 6. Number of Instances in Which Results of State Assessments Were Comparable to State NAEP Results (Total Possible Instances = 60)

	Number of Instances in Which Results Were Similar or Different	
Achievement Level	**Similar**	**Different**
Advanced	7	8
Proficient	12	3
Basic	9	6
Below Basic	5	10
All Levels	33	27

It should be noted that a larger sample of individual State and State NAEP results should be collected and analyzed before any definitive conclusions can be reached regarding the extent to which the State and NAEP assessments results agree. However, the modest consistency in the results reported by State assessments and by State NAEP results shown in this section suggests that many States have not only established performance standards that are similar to NAEP, but have reported assessment results that are similar to NAEP as well.[1]

[1] Editor's Note: A different picture would emerge and different conclusions might be drawn if the percentages in Tables 7–15 were displayed as cumulative percentages rather than as discreet percentages within each performance level category. For example, in Table 7, for two comparable levels like Partially Proficient (CO) and Basic (NAEP), CO has 87% at or above Partially Proficient while NAEP has 69% at or above Basic. Comparing the within-category percentages, the difference is a more modest 5%.

Table 7. Colorado: Bookmark Standard-Setting Method

State Assessment Results: Colorado		**State NAEP Results: Colorado**	
Achievement Level	Percentage of Students Grade 4 Reading 1998	Achievement Level	Percentage of Students Grade 4 Reading 1998
Advanced	6	Advanced	7
Proficient	51	Proficient	27
Partially Proficient	30	Basic	35
Unsatisfactory	10	Below Basic	31
Not Tested	3		

Shading is used to show when the results of the two assessments at a particular achievement level are similar.

Table 8. Kentucky: Booklet Classification Standard-Setting Method

State Assessment Results: Kentucky			**State NAEP Results: Kentucky**		
Achievement Level	Percentage of Students		Achievement Level	Percentage of Students	
	Grade 4 Reading 1998	Grade 8 Math 1998		Grade 4 Math 1996	Grade 8 Math 1996
Distinguished	12	14	Advanced	1	1
Proficient	10	19	Proficient	15	15
Apprentice	55	36	Basic	44	40
Novice	23	29	Below Basic	40	44

Shading is used to show when the results of the two assessments at a particular achievement level are similar.

Table 9. Louisiana: Bookmark Standard-Setting Method

State Assessment Results: Louisiana		State NAEP Results: Louisiana	
Achievement Level	Percentage of Students Grade 4 Math 1999	Achievement Level	Percentage of Students Grade 4 Math 1996
Advanced	2	Advanced	0
Proficient	8	Proficient	8
Basic	32	Basic	36
Approaching Basic	36	Below Basic	56
Unsatisfactory	21		

Shading is used to show when the results of the two assessments at a particular achievement level are similar.

Table 10. Maine: Booklet Classification Standard-Setting Method

State Assessment Results: Maine			State NAEP Results: Maine		
Achievement Level	Percentage of Students		Achievement Level	Percentage of Students	
	Grade 4 Reading 1998	Grade 4 Math 1998		Grade 4 Reading 1998	Grade 4 Math 1996
Distinguished	1	7	Advanced	8	3
Advanced	22	13	Proficient	28	24
Basic	66	52	Basic	37	48
Novice	11	28	Below Basic	27	25

Shading is used to show when the results of the two assessments at a particular achievement level are similar.

Table 11. Massachusetts: Booklet Classification Standard-Setting Method

State Assessment Results: Massachusetts			**State NAEP Results: Massachusetts**		
Achievement Level	Percentage of Students		Achievement Level	Percentage of Students	
	Grade 4 Math 1998	Grade 8 Math 1998		Grade 4 Math 1996	Grade 8 Math 1996
Advanced	11	8	Advanced	2	5
Proficient	23	23	Proficient	22	23
Needs Improvement	44	26	Basic	47	40
Failing	23	42	Below Basic	29	32

Shading is used to show when the results of the two assessments at a particular achievement level are similar.

Table 12. Minnesota: Modified Angoff Standard-Setting Method

State Assessment Results: Minnesota			**State NAEP Results: Minnesota**		
Achievement Level	Percentage of Students		Achievement Level	Percentage of Students	
	Grade 5 Reading 1999	Grade 5 Math 1999		Grade 4 Reading 1998	Grade 4 Math 1996
Level IV	12	6	Advanced	8	3
Level III	33	31	Proficient	28	26
Level II	37	45	Basic	33	47
Level I	18	18	Below Basic	31	24

Shading is used to show when the results of the two assessments at a particular achievement level are similar.

Table 13. Missouri: Bookmark Standard-Setting Method

State Assessment Results: Missouri		State NAEP Results: Missouri	
Achievement Level	Percentage of Students Grade 4 Math 1999	Achievement Level	Percentage of Students Grade 4 Math 1996
Advanced	6	Advanced	1
Proficient	29	Proficient	20
Nearing Proficient	43	Basic	46
Progressing	19	Below Basic	34
Step 1	3		

Shading is used to show when the results of the two assessments at a particular achievement level are similar.

Table 14. New Mexico: Bookmark Standard-Setting Method

State Assessment Results: New Mexico			State NAEP Results: New Mexico		
Achievement Level	Percentage of Students		Achievement Level	Percentage of Students	
	Grade 4 Math 1999	Grade 8 Math 1999		Grade 4 Math 1996	Grade 8 Math 1996
Advanced	12	6	Advanced	1	2
Proficient	20	18	Proficient	12	12
Nearing Proficiency	48	27	Basic	38	37
Beginning Step	19	50	Below Basic	49	49

Shading is used to show when the results of the two assessments at a particular achievement level are similar.

Table 15. Washington: Bookmark Standard-Setting Method

State Assessment Results: Washington			**State NAEP Results: Washington**		
Achievement Level	Percentage of Students		Achievement Level	Percentage of Students	
	Grade 4 Reading 1998	Grade 4 Math 1997		Grade 4 Reading 1998	Grade 4 Math 1996
Level IV	16	7	Advanced	6	1
Level III	40	15	Proficient	23	21
Level II	35	29	Basic	34	46
Level I	8	47	Below Basic	37	33
Not Tested	2	3			

Shading is used to show when the results of the two assessments at a particular achievement level are similar.

Conclusion

The findings of this study provide a State-level context for NAGB to consider as it discusses refinements to the performance standards currently used to report the results of NAEP. First, although it can be argued that few, if any, States have developed performance standards that are absolutely equivalent to NAEP's standards, many States appear to have borrowed and learned from NAEP to report the results of their assessment programs. An analysis of the component parts of State performance standards indicates that at least 23 States have developed achievement level categories that are similar to NAEP. Moreover, when surveyed about whether they perceive their state's performance standards to be similar to NAEP, 17 of 34 State testing directors responded affirmatively. Although most States have adopted different names for their achievement level categories, Arkansas, California, Pennsylvania, South Carolina, and South Dakota have adopted (or plan to adopt) the same names—Advanced, Proficient, Basic, and Below Basic.

Unlike NAEP, at least eight States have adopted five rather than four achievement levels. In those States, it appears they have divided what is roughly equivalent to NAEP's Below Basic level into two parts. Although it was beyond the scope of this study to determine why States decided to do this, it seems reasonable to speculate that because relatively large percentages of students continue to perform at the Below Basic level, that States (1) need to identify students at the low end of the Below Basic range for targeting additional resources, and (2) need a reporting system that is sensitive to changes over time within the Below Basic category. Some have also argued that the Basic level be subdivided for similar reasons.

Second, this study shows that most States have used one of three methods to set standards: the modified Angoff method, the bookmark method, or the booklet classification method. The Bookmark method, used in 18 States, was by far the most commonly used. Most States that use the bookmark method contract with CTB/McGraw Hill to assist in implementing their state's assessment program. Similarly, States using the booklet classification method tend to contract with Advanced Systems, which employs a version of the booklet classification method they refer to as the body of work method, while States using the modified Angoff method tend to contract with Harcourt Brace. Further investigation is needed to determine whether States actively evaluated standard-setting options before selecting one or whether they settled for the method preferred by their respective testing contractor.

Although the final part of this study was based on a relatively small sample of cases, the findings suggest that there may be a moderate concurrence in the results reported by State NAEP and comparable State assessment programs, which have adopted NAEP-like performance standards. The results of State NAEP and State assessment programs tend to be most consistent at the middle two levels (Proficient and Basic). In general, States tend to report a higher percentage of students at the Advanced level and a lower percentage of students at the bottom level than State NAEP. Further investigation is required to explain whether this is the result of differences in the standard-setting methods, the use of impact data in setting standards, the descriptions of the achievement level categories, test content, the policy environment in which the standards were set, or other as yet undetermined factors.

References

List of State Performance Standards

Alabama Department of Education. (1999). *Alabama direct assessment of writing: Grade five focused holistic rubric*. Montgomery, AL.

Alaska Department of Education. (1999 Draft). *Alaska writing reading standards*. Juneau, AK.

Arizona Department of Education. (1999). *State Board Approved AIMS Performance Levels Grades K–12.* Phoenix, AZ.

Arkansas Department of Education. (1999). *Arkansas general performance level definitions.* Little Rock, AR.

California Department of Education. (1999). *English language arts proposed descriptors for performance standards/levels Grade 5.* Sacramento, CA.

Colorado Department of Education. (1998). *Colorado student assessment program.* Denver, CO.

Connecticut Department of Education. Connecticut (1998a). *Academic performance test 1998 CAPT program overview.* Hartford, CT.

Connecticut Department of Education. (1999b). *Connecticut mastery test 1999 CMT program overview.* Hartford, CT.

Council of Chief State School Officers. (1996a). *Annual survey of State student assessment programs, Fall 1998.* Washington, DC.

Council of Chief State School Officers. (1997b). *Annual Survey of State Student Assessment Programs, Fall 1997.* Washington, DC.

Council of Chief State School Officers. (1998c). *Annual Survey of State Student Assessment Programs, Fall 1998.* Washington, DC.

Delaware Department of Public Instruction. (1999). *Delaware Student Testing Program 1999 Administration State Summary Report.* Dover, DE.

Donahue, P. L., Voekl, K. E., Campbell, J. R., and Mazzeo, J. (1999). *The NAEP 1998 reading report card for the Nation and the States.* NCES 1999–500. Washington, D.C.: National Center for Education Statistics.

Florida Department of Education. (1999). *Performance levels set.* Tallahassee, FL.

Idaho Department of Education. (1999). *Idaho direct mathematics assessment scoring standard.* Boise, ID.

Illinois State Board of Education. (1999). *Understanding your child's ISAT scores.* Springfield, IL.

Kentucky Department of Education. (1999 Draft). *Descriptors of four-point rubric.* Frankfort, KY.

Louisiana Department of Education. (1999). *LEAP for the 21st century, spring 1999 Criterion-Referenced Test (CRT) State/district level summary report.* Baton Rouge, LA.

Maine Department of Education. (1998). *Maine educational assessment system State summary scores.* Augusta. ME.

Maryland Department of Education. (1998). *Definitions performance standards.* Baltimore, MD.

Massachusetts Department of Education. (1998). *The Massachusetts comprehensive assessment system: Summary of district performance.* Malden, MA.

Michigan Department of Education. (1999). *Performance criteria: High school tests.* Lansing, MI.

Minnesota Department of Children, Families and Learning. (1999). *State profile report 1998/1999.* St. Paul, MN.

Missouri Department of Education. (1999). *State results.* Jefferson City, MO.

New Hampshire Department of Education. (1999). *Proficiency level definitions.* Concord, NH.

New Jersey Department of Education (1998). *New Jersey statewide testing system March 1998: Grade 8 early warning test reading, mathematics, and writing State results.* Trenton, NJ.

New Mexico Department of Education. (1999). *Statewide articulated assessment system 1998–1999 summary report.* Santa Fe, NM.

New York Department of Education. (1999). *Understanding performance levels for Grade 4 English language arts.* Albany, NY.

Oklahoma Department of Education. (1999). *Oklahoma writing assessment summary report.* Oklahoma City, OK.

Reese, C. M., Miller, K. E., Mazzeo, J., and Dossey, J. A. (1997). *The NAEP 1996 mathematics report card for the Nation and the States.* Washington, D.C.: National Center for Education Statistics.

Rhode Island Department of Education. (1999). *English language arts reference examination State summary for Rhode Island.* Providence, RI.

Tennessee State Department of Education. (1998). *21st century report card Tennessee value-added assessment system.* Nashville, TN.

Virginia Department of Education. (1999). *Virginia standards of learning assessments statewide passing rates: spring 1998 and spring 1999.* Richmond, VA.

Washington Office of the Superintendent of Public Instruction. (1999). *Summary of student performance: 4th, 7th, and 10th grades.* Olympia, WA.

Wisconsin Department of Public Instruction. (1997). *Final summary report of the proficiency score standards.* Madison, WI.

Wyoming Department of Education. (1999). *1999 WyCAS State level results.* Cheyenne, WY.

Appendix A: Achievement Level Reporting Categories in State Assessment Programs

State	Assessment Program*	Achievement Level Reporting Categories	Method Used To Set Cut-Scores	Similarity of State Achievement Levels and Standard-Setting Method to NAEP (Same, Similar, Different)				Overall Similarity of State Performance Standards to NAEP
				Number of Levels	Name of Each Level	Definition of Levels	Method Used To Set Cut-Scores	
Alabama	Writing in grades 5 and 7	Level IV Level III Level II Level I	Levels correspond directly to scoring guide	Same	Different	Similar	Different	**Different**
Alaska	Reading, writing, and math in grades 3, 6, 8, and 10 (beginning in 1999–2000)	Advanced Proficient Below Proficient Not Proficient	Will use bookmark	Same	Similar	Similar	Similar	**Similar**
Arizona	Reading, writing, and math in grades 3, 5, and 8 (administered for the first time in 1998–1999)	Exceeds the Standard Meets the Standard Approaches the Standard Falls Far Below the Standard	Bookmark	Same	Similar	Similar	Similar	**Similar**
Arkansas	Reading, writing, and math in grade 4	Advanced Proficient Basic Below Basic	Booklet classification	Same	Same	Similar	Similar	**Similar**
California	Language arts, math, science, and history in grade 2–11 (beginning 1999–2000)	Advanced Proficient Basic Below Basic	Modified Angoff	Same	Same	Insufficient information	Same	**Insufficient information**
Colorado	Reading in grades 3 and 4; writing in grade 4	Advanced Proficient Partially Proficient Unsatisfactory	Bookmark	Same	Similar	Similar	Similar	**Similar**

* This table includes assessment programs for which states use achievement level categories to report results only. Consequently, states may administer other assessments than those listed. Unless noted otherwise, the assessment programs listed were administered during the 1997–1998 school year.

Appendix A: Achievement Level Reporting Categories in State Assessment Programs

State	Assessment Program*	Achievement Level Reporting Categories	Method Used To Set Cut-Scores	Similarity of State Achievement Levels and Standard-Setting Method to NAEP (Same, Similar, Different)				Overall Similarity of State Performance Standards to NAEP
				Number of Levels	Name of Each Level	Definition of Levels	Method Used To Set Cut-Scores	
Connecticut	Reading in grades 4, 6, and 8	State Goal Below State Goal Well Below State Goal	Modified Angoff	Different	Different	Different	Same	**Different**
	Math in grades 4, 6, and 8	State Goal Slightly Below State Goal Below State Goal Well Below State Goal	Modified Angoff	Same	Different	Different	Same	**Different**
	Writing in grades 4, 6, and 8	Well Above State Goal State Goal Slightly Below State Goal Well Below State Goal	Modified Angoff	Same	Different	Different	Same	**Different**
	Math, science, response to literature, interdisciplinary in grade 10	State Goal Somewhat Below State Goal Below State Goal Well Below State Goal	Modified Angoff	Same	Different	Different	Same	**Different**
	Editing in grade 10	State Standard Below State Standard	Modified Angoff	Different	Different	Different	Same	**Different**
Delaware	Reading, writing, and math in grades 3, 5, 8, and 10 (first administered in 1998–1999)	Distinguished Exceeds Standard Meets Standard Below Standard Well Below Standard	Bookmark	Different	Similar	Similar	Similar	**Similar**

* This table includes assessment programs for which states use achievement level categories to report results only. Consequently, states may administer other assessments than those listed. Unless noted otherwise, the assessment programs listed were administered during the 1997–1998 school year.

Appendix A: Achievement Level Reporting Categories in State Assessment Programs

State	Assessment Program*	Achievement Level Reporting Categories	Method Used To Set Cut-Scores	Similarity of State Achievement Levels and Standard-Setting Method to NAEP (Same, Similar, Different)				Overall Similarity of State Performance Standards to NAEP
				Number of Levels	Name of Each Level	Definition of Levels	Method Used To Set Cut-Scores	
Florida	Writing in grades 4, 8, and 10	Level 5 Level 4 Level 3 Level 2 Level 1	Levels directly correspond with scoring rubric levels	Different	Different	Different	Different	**Different**
	Reading in grades 4, 8, and 10; math in grades 5, 8, and 10	Levels 1–5 (1 is the lowest level)	Multistep process by 90-person study committee	Different	Different	Different	Different	**Different**
Georgia	Writing in grade 8	Very Good Good Minimal Inadequate	Based on scoring criteria	Same	Different	Different	Different	**Different**
Hawaii	15 essential competencies in grades 10–12	Pass Fail	Minimum performance standards determined using chi-square distribution method	Different	Different	Different	Different	**Different**
Idaho	Math in grades 4 and 8; writing in grades 4, 8, and 11	Advanced Proficient Satisfactory Developing Minimal	Committees of teachers and Idaho Department of Education staff	Different	Similar	Different	Different	**Different**

* This table includes assessment programs for which states use achievement level categories to report results only. Consequently, states may administer other assessments than those listed. Unless noted otherwise, the assessment programs listed were administered during the 1997–1998 school year.

Appendix A: Achievement Level Reporting Categories in State Assessment Programs

State	Assessment Program*	Achievement Level Reporting Categories	Method Used To Set Cut-Scores	Similarity of State Achievement Levels and Standard-Setting Method to NAEP (Same, Similar, Different)				Overall Similarity of State Performance Standards to NAEP
				Number of Levels	Name of Each Level	Definition of Levels	Method Used To Set Cut-Scores	
Illinois	Reading, writing, and math in grades 3, 6, 8, and 10	Exceeds Standards Meets Standards Below Standards Academic Warning	Modified Angoff	Same	Different	Similar	Same	**Similar**
Indiana	Language arts and math in grades 3, 6, 8, and 10	Above Standards Below Standards	Bookmark	Different	Different	Different	Similar	**Different**
Iowa	Reading and math in grades 4, 8, and 11	High Intermediate Low	Insufficient information	Different	Different	Different	In sufficient information	**Different**
Kansas	Reading in grades 3, 7, and 10; writing in grades 5, 8, and 10; math in grades 4, 7, and 10	Excellent Proficient Basic Unsatisfactory	Performance distribution estimators	Similar	Similar	Insufficient information	Different	**Different**
Kentucky	Reading, writing, and science in grades 4, 7, and 11; math, social studies, arts and humanities, and practical living in grades 5, 8, and 11	Distinguished Proficient Apprentice Novice	Booklet classification	Same	Similar	Similar	Similar	**Similar**
Louisiana	English and math in grades 4 and 8 in spring 1999 and in grade 10 in 2001; science and social studies in grades 4 and 8 in 2000 and in grade 11 in 2002	Advanced Proficient Basic Approaching Basic Unsatisfactory	Bookmark	Different	Similar	Similar	Similar	**Similar**

* This table includes assessment programs for which states use achievement level categories to report results only. Consequently, states may administer other assessments than those listed. Unless noted otherwise, the assessment programs listed were administered during the 1997–1998 school year.

Appendix A: Achievement Level Reporting Categories in State Assessment Programs

State	Assessment Program*	Achievement Level Reporting Categories	Method Used To Set Cut-Scores	Similarity of State Achievement Levels and Standard-Setting Method to NAEP (Same, Similar, Different)				Overall Similarity of State Performance Standards to NAEP
				Number of Levels	Name of Each Level	Definition of Levels	Method Used To Set Cut-Scores	
Maine	Reading, writing, math, science/technology, and social studies	Exceeds Standards Meets Standards Partially Meets Standards Does Not Meet Standards	Booklet classification	Same	Similar	Similar	Similar	**Similar**
Maryland	Reading, writing, language usage, math, science, and social studies in grades 3, 5, and 8	Level 1 – Excellent Level 2 – Excellent Level 3 – Satisfactory Level 4 – Not Met Level 5 – Not Met	Delphi method including impact data	Different	Different	Different	Different	**Different**
Massachusetts	English language arts, math, science/technology, and history/social science in grades 4, 8, and 10	Advanced Proficient Needs Improvement Failing	Booklet classification	Same	Similar	Similar	Similar	**Similar**
Michigan	Reading and math in grades 4 and 7	Satisfactory, Moderate, Low	Modified Angoff	Different	Different	Different	Same	**Different**
	Writing in grades 5 and 8	Proficient, Not Yet Proficient	Modified Angoff and contrasting groups	Different	Different	Different	Similar	**Different**
	Science in grades 5 and 8	Proficient, Novice, Not Yet Novice	Modified Angoff and contrasting groups	Different	Different	Different	Similar	**Different**
	Reading, writing, math, and science in grade 11 (1998)	Level 1 – Exceeds Level 2 – Met Level 3 – Basic Level 4 – Did not meet	Bookmark	Same	Similar	Similar	Similar	**Similar**

* This table includes assessment programs for which states use achievement level categories to report results only. Consequently, states may administer other assessments than those listed. Unless noted otherwise, the assessment programs listed were administered during the 1997–1998 school year.

Appendix A: Achievement Level Reporting Categories in State Assessment Programs

State	Assessment Program*	Achievement Level Reporting Categories	Method Used To Set Cut-Scores	Similarity of State Achievement Levels and Standard-Setting Method to NAEP (Same, Similar, Different)				Overall Similarity of State Performance Standards to NAEP
				Number of Levels	Name of Each Level	Definition of Levels	Method Used To Set Cut-Scores	
Minnesota	Reading and math in grades 3, 4, and 5; writing in grade 5	Level IV Level III Level II Level I	Modified Angoff	Same	Different	Similar	Same	**Similar**
Mississippi	Reading, written communication, and math in grade 11	Pass Fail	Insufficient information	Different	Different	Different	Insufficient information	**Different**
Missouri	Language arts and math in grades 2–10; science and social studies in grades 3–10	Advanced Proficient Nearing Proficient Progressing Step 1	Bookmark	Different	Similar	Similar	Similar	**Similar**
Montana	Reading, language arts, math, science, and social studies in grades 4, 8, and 11	Advanced Proficient Nearing Proficiency Novice	Grouped stanine scores: Novice 1–3, Nearing 4, Proficient 5–7, Advanced 8–9	Same	Similar	Different	Different	**Different**
Nebraska	Plans to implement a statewide assessment program in 1999–2000							
Nevada	Writing in grade 8	Level 5 Level 4 Level 3 Level 2 Level 1	Based on scoring rubric	Different	Different	Different	Different	**Different**

* This table includes assessment programs for which states use achievement level categories to report results only. Consequently, states may administer other assessments than those listed. Unless noted otherwise, the assessment programs listed were administered during the 1997–1998 school year.

Appendix A: Achievement Level Reporting Categories in State Assessment Programs

State	Assessment Program*	Achievement Level Reporting Categories	Method Used To Set Cut-Scores	Similarity of State Achievement Levels and Standard-Setting Method to NAEP (Same, Similar, Different)				Overall Similarity of State Performance Standards to NAEP
				Number of Levels	Name of Each Level	Definition of Levels	Method Used To Set Cut-Scores	
New Hampshire	English language arts and math in grades 3, 6, and 10; science and social studies in grades 6 and 10	Advanced Proficient Basic Novice	Booklet classification	Same	Similar	Similar	Similar	**Similar**
New Jersey	Language arts, math, and science in grade 4; reading, writing, and math in grade 8	Level 1 (High) Level 2 (Minimal) Level 3 (Below Minimal)	Booklet classification (holistic portfolio)	Different	Different	Different	Similar	**Different**
New Mexico	Reading, language arts, science and social studies in grades 4, 6, and 8	Advanced Proficient Nearing Proficiency Beginning Step	Bookmark	Same	Similar	Similar	Similar	**Similar**
New York	English language arts and math in grades 4 and 8	Level 4 Level 3 Level 2 Level 1	Item mapping, checking against operational data	Same	Different	Similar	Different	**Different**
North Carolina	Reading and math in grades 3–9, writing in grades 4, 7, and 10	Level IV Level III Level II Level I	Contrasting groups	Same	Different	Different	Different	**Different**
North Dakota	Reading, language arts, math, science, and social studies in grades 4, 6, 8, and 10	Advanced Proficient Partially Proficient Novice	Based on national percentile rankings	Same	Similar	Insufficient information	Different	**Different**

* This table includes assessment programs for which states use achievement level categories to report results only. Consequently, states may administer other assessments than those listed. Unless noted otherwise, the assessment programs listed were administered during the 1997–1998 school year.

Appendix A: Achievement Level Reporting Categories in State Assessment Programs

State	Assessment Program*	Achievement Level Reporting Categories	Method Used To Set Cut-Scores	Similarity of State Achievement Levels and Standard-Setting Method to NAEP (Same, Similar, Different)				Overall Similarity of State Performance Standards to NAEP
				Number of Levels	Name of Each Level	Definition of Levels	Method Used To Set Cut-Scores	
Ohio	Reading and writing in grades 4 and 6 Reading and writing in grade 12	Advanced Proficient Partially Proficient Honors Proficient Below Proficient	Committee of educators	Different	Similar	Different	Different	**Different**
Oklahoma	Reading, writing, math, science, geography, and history in grades 5, 8, and 11	Advanced Satisfactory Limited Knowledge Unsatisfactory	Bookmark	Same	Similar	Similar	Similar	**Similar**
Oregon	Reading and math in grades 3, 5, 8, and 10; writing in grades 5, 8, and 10	Exceeds Standards Meets Standards Does Not Yet Meet Standards	Bookmark and iterative review of student work	Different	Different	Insufficient information	Different	**Different**
Pennsylvania	Reading and math in grades 5, 8, and 11; writing in grades 3, 7, and 10	Advanced Proficient Basic Below Basic (beginning in 2001)	Plans to use bookmark	Same	Same	Similar	Similar	**Will be similar**
Puerto Rico	English, math, science, social studies, and Spanish in grades 3, 6, 9, and 11	Highly Competent Competent Partially Competent	Insufficient information	Different	Different	Insufficient information	Insufficient information	**Different**

* This table includes assessment programs for which states use achievement level categories to report results only. Consequently, states may administer other assessments than those listed. Unless noted otherwise, the assessment programs listed were administered during the 1997–1998 school year.

Appendix A: Achievement Level Reporting Categories in State Assessment Programs

State	Assessment Program*	Achievement Level Reporting Categories	Method Used To Set Cut-Scores	Similarity of State Achievement Levels and Standard-Setting Method to NAEP (Same, Similar, Different)				Overall Similarity of State Performance Standards to NAEP
				Number of Levels	Name of Each Level	Definition of Levels	Method Used To Set Cut-Scores	
Rhode Island	English language arts (reading and writing) and math in grades 4, 8 and 10	Achieved Standard with Honors Achieved Standard Nearly Achieved Standard Below Standard Little Evidence of Achievement	Review of profiles of student work	Different	Similar	Similar	Similar	**Similar**
	Writing in grades 3, 7, and 10; Health in grades 5 and 9	Same as above	Modified Angoff (and impact data)	Different	Similar	Similar	Different for writing	**Different for writing**
South Carolina	Basic Skills Assessment Program in grades 3, 6, 8, and 10	Currently Pass/Fail In future will use NAEP categories' names only	Not yet identified	Different	Will be similar	Will be different	Not yet identified	**Will be different**
South Dakota	Reading, language arts, math, science, and social studies	Advanced Proficient Basic Below Basic	Angoff	Same	Same	Insufficient information	Similar	**Insufficient information**
Tennessee	Reading, language arts, math, science, and social studies in grades 3, 4, 5, 6, 7, and 8	Advanced Proficient Nearly Proficient Progressing Step 1	Bookmark	Different	Similar	Similar	Similar	**Similar**
	Writing in grades 4, 7, and 11	6 – Outstanding 5 – Strong 4 – Competent 3 – Limited 2 – Flawed 1 – Flawed/Deficient	Cut points correspond to scoring rubric	Different	Different	Different	Different	**Different**

* This table includes assessment programs for which states use achievement level categories to report results only. Consequently, states may administer other assessments than those listed. Unless noted otherwise, the assessment programs listed were administered during the 1997–1998 school year.

Appendix A: Achievement Level Reporting Categories in State Assessment Programs

State	Assessment Program*	Achievement Level Reporting Categories	Method Used To Set Cut-Scores	Similarity of State Achievement Levels and Standard-Setting Method to NAEP (Same, Similar, Different)				Overall Similarity of State Performance Standards to NAEP
				Number of Levels	Name of Each Level	Definition of Levels	Method Used To Set Cut-Scores	
Texas	Reading and math in grades 3–8 and exit level; writing at grades 4 and 8 and exit level; science and social studies at grade 8; end-of-course exams also in algebra I and biology I (1996–1997)	Minimum Expectations (70% correct) Mastered All Objectives Academic Recognition (95% correct)	Set by board of education	Different	Different	Different	Different	**Different**
U.S. Virgin Islands	Reading, language arts, math, science, and social studies in grades 3, 6, 8 and 11	Advanced Proficient Basic Below Basic	Insufficient information	Same	Similar	Insufficient information	Insufficient information	**Insufficient information**
Utah	Language arts, math, and science tests under development	Not yet established	May use bookmark					
Vermont	English language arts and math in grades 4, 8, and 10; science in grades 6 and 11	Honors Standard Nearly Standard Below Standard Little Evidence of Achievement	Science – bookmark	Different	Similar	Similar	Similar	**Similar**

* This table includes assessment programs for which states use achievement level categories to report results only. Consequently, states may administer other assessments than those listed. Unless noted otherwise, the assessment programs listed were administered during the 1997–1998 school year.

Appendix A: Achievement Level Reporting Categories in State Assessment Programs

State	Assessment Program*	Achievement Level Reporting Categories	Method Used To Set Cut-Scores	Similarity of State Achievement Levels and Standard-Setting Method to NAEP (Same, Similar, Different)				Overall Similarity of State Performance Standards to NAEP
				Number of Levels	Name of Each Level	Definition of Levels	Method Used To Set Cut-Scores	
Virginia	English, math, history, and social science in grades 3, 5, and 8; computer/technology in grade 8; English, algebra I and II, geometry, earth science, biology, chemistry, world history, and U.S. history in grade 10	Advanced (High) Proficient (Passing) Does Not Meet (Failing)	Modified Angoff	Different	Similar	Different	Same	**Different**
Washington	Reading and math in grades 4 and 7	Level 4 (Above Standard) Level 3 (Meets Standard) Level 2 (Below Standard) Level 1 (Well Below Standard)	Bookmark	Same	Similar	Similar	Similar	**Similar**
West Virginia	Reading, language arts, math, science, and social studies in grades 3–11 Writing in grades 4, 7, and 10	Under development						
Wisconsin	Reading, writing, math, science, and social studies in grades 4, 8 and 10	Advanced Proficient Basic Minimal	Bookmark	Same	Similar	Similar	Similar	**Similar**
Wyoming	Reading, writing, and math in grades 4, 8, and 11	Advanced Proficient Partially Proficient Novice	Booklet classification	Same	Similar	Similar	Similar	**Similar**

* This table includes assessment programs for which states use achievement level categories to report results only. Consequently, states may administer other assessments than those listed. Unless noted otherwise, the assessment programs listed were administered during the 1997–1998 school year.

SECTION 6

Newspaper Coverage of NAEP Results, 1990 to 1998

Ronald K. Hambleton and Kevin Meara

University of Massachusetts at Amherst

November 2000

Newspaper Coverage of NAEP Results, 1990 to 1998[1, 2]

Ronald K. Hambleton and Kevin Meara

Background and Purposes

Since the inception of the National Assessment of Educational Progress (NAEP) in the 1960s, questions have been raised about the extent to which policymakers and the public understand NAEP scores and their usefulness (see, e.g., Barron and Koretz, 1998; Hambleton and Slater, 1994; Jaeger, 1998; Koretz and Deibert, 1993; Messick, Beaton, and Lord, 1983; Wainer, Hambleton, and Meara, 1999). The shift to an item response model-based reporting system in 1984 from an exercise-by-exercise reporting system used previously was one major initiative taken by the U.S. Department of Education to improve the understandability of NAEP score reporting. The most important reason for the National Assessment Governing Board (NAGB) introduction of performance standards, or achievement levels (as they are called by NAGB), beginning with the 1990 NAEP Mathematics Assessment was to increase the understandability and usefulness of NAEP results for policymakers, educators, and the public.

Several studies have investigated NAEP reporting to policymakers, educators and the public in the 1990s (Hambleton and Slater, 1994; Wainer, Hambleton, and Meara, 1999). Now, 10 years after the introduction of achievement levels in NAEP score reporting, there is merit to reviewing the way newspapers have been presenting and interpreting NAEP results to the public. How central are the achievement levels in newspaper reports of NAEP results? What other NAEP information do the newspapers report, and how well are they reporting it? These were the types of questions addressed in this research study. Specifically, this study was designed to answer the following:

1. How have NAEP press briefing packages changed over the past 10 years?
2. What information has been highlighted in the newspaper accounts of NAEP results?
3. Is there evidence that NAEP press release materials are being understood and used by the newspapers in their stories?
4. Are the newspapers accurately conveying information about NAEP results to their readers?

The study by Barron and Koretz (1998) covered some of the same questions as this study, but their work is different from the current study in several important ways. This study is focused on reporting national results and the Barron and Koretz study was focused on reporting trial State assessment results; press release packages were considered in this study, but not in the Barron and Koretz study; and this research study includes information about the recent releases of the

[1] *Laboratory of Psychometric and Evaluative Research Report No. 366*. Amherst, MA: University of Massachusetts, School of Education.

[2] Financial support for conducting this research study was provided by the National Assessment Governing Board (NAGB). However, opinions expressed in the article are those of the researchers and no NAGB endorsement should be assumed.

1998 reading and writing assessments, and the Barron and Koretz study did not. The Barron and Koretz study was focused on only two NAEP releases: 1994 Reading and 1996 Mathematics. This study was a less detailed review of NAEP releases of national results than the Barron and Koretz study was of the trial State assessment results, but findings from 7 of the 10 NAEP national releases of results using achievement levels since 1990 were reviewed, including two of the three releases in 1998 when more focus was placed on NAEP score reporting than in previous releases. Time limitations prevented a review of the 1994 Geography and U.S. History results and 1998 Civics results.

Study Design

NAGB contracted with Bacon's Electronic Clipping Service to select newspaper stories about NAEP releases of national results. Clippings were from stories published within 2 or 3 weeks of NAEP score releases. More than 500 clippings from NAEP releases in the past decade were reviewed. The major drawback of the clipping service was that graphics were not included. Often the graphics are a major source of the problem in reporting NAEP results (Hambleton and Slater, 1994). Keywords used in the search of more than 140 U.S. newspapers included National Assessment of Educational Progress, NAEP, achievement levels, student standards, subject areas of the NAEP reports (mathematics, reading, science, and writing), and specific dates of the releases. In addition, we were able to obtain the press release documents for nearly all of the NAEP releases since 1990. These included Mathematics, 1990, 1992, 1996; Reading, 1994, 1998; Science, 1996; and Writing, 1998.

A copy of the coding form that was designed for cataloging the newspaper stories is presented in appendix A. Records were kept on variables such as the year, subject, article title, newspaper, publication date, newspaper size, length of article, and presence of graphics. In addition, the articles were reviewed for 16 features: discussion of the standards, reporting of scaled scores, national results, State-by-State information, State to national information, changes over time, curriculum consideration, NAEP over test comparisons, NAEP limitations, multiple-subject reporting, interesting anecdotes or examples, and reporting of sex, race, socioeconomic status (SES), parent, and interaction information.

Results and Discussion

Results are reported in eight sections. First, a comparison of press releases for mathematics in 1990, 1992, and 1996 is presented. Mathematics was the only subject area that was assessed three times in the past decade. Next, an analysis of newspaper clippings for mathematics in the 1992 and 1996 releases is presented. Reading was assessed twice in the 1990s with achievement levels. A comparison of reading press releases for 1994 and 1998 is made. An analysis of news clippings for 1994 and 1998 follows. This pattern in presenting results is repeated for 1996 Science and 1998 Writing.

Comparison of Press Releases for Mathematics: 1990, 1992, and 1996

1990 Mathematics. The 1990 Mathematics press release served as a baseline for this research, because that was the first year student performance was reported in terms of achievement levels. Previously, results were presented in terms of NAEP scaled scores. The main focus of the 1990 Mathematics press release materials was NAGB's new standards or achievement levels. The data presentation indicated that "the Board's new standards allow NAEP data to be reported in terms of what students *should* be able to do." Immediately, in the 2-page executive summary report, the achievement levels were introduced and defined as follows:

The Basic level denotes partial mastery of the knowledge and skills fundamental for Proficient work at each grade. Proficient, the central level, represents solid academic performance and demonstrated competence over challenging subject matter. The Advanced level signifies superior performance beyond Proficient.

Unfortunately, the 1990 national results were disappointing, and the report indicated that most students were not demonstrating the performance required to meet the Proficient level. In our opinion, despite the best efforts of NCES and NAGB at the time, the press might have had difficulties understanding what the achievement levels were about. The definitions and statements about the achievement levels appeared clear and straightforward to us, but they were not necessarily sufficient for policymakers and the public to grasp. In addition, some of the wording may have led to confusion. For example, the phrase "less than 20% reached Proficient levels" does not clearly indicate whether students in the Advanced category are included as part of the 20%. Of course, as it turned out, the percentage of students at the Advanced level was small, but this would not have been known to readers at the time.

National averages were reported in terms of achievement levels (e.g., just over 60% of students in 4th, 8th, and 12th grades were at or above Basic, less than 20% reached Proficient, and only 0.6 to 2.6% of the students reached the Advanced level). Scores were also reported in relation to several variables, including sex, race, type of community (SES), parental education, and number of mathematics courses taken (grade 12 only). The performance of public versus private school students was not reported. No graphics appear to have been used with this 1990 press release, and average scaled scores were not reported.

The central message of the 1991 press conference was that this new way of reporting achievement was important and useful; unfortunately, the results were disappointing. In addition, because there had been criticism of the new achievement levels (and how they were set), some statements were made defending the achievement levels as "strong" and "set by a reasonable process."

1992 Mathematics. The press release materials for the 1992 Mathematics report had a different feel than the 1990 report. Less emphasis was placed on the achievement levels, although they were defined and discussed. Examples of test items were presented to help the press understand what students at each achievement level were expected to be able to do. The 1992 materials presented the national results for 4th, 8th, and 12th graders but also presented results for 4th and

8th graders in 44 States and territories. Comparisons were made between 1990 and 1992 performance for both the Nation and for 8th graders in 37 jurisdictions. Also new in 1992 was the use of the word cut-point to describe the standards or achievement levels. Caution was expressed that the "National Center for Education Statistics (NCES) lacked compelling evidence about what inferences could be drawn from the NAEP results." NCES was not convinced that students who were Proficient could actually do all the things covered in the achievement-level descriptions. This concern was addressed in subsequent NAEP releases by highlighting the capabilities of students who were placed in the performance categories and other students who performed exactly at the achievement levels.

Figures were used to present the results at the press conference and in a report, but we did not have access to these figures. Also new in 1992 was use of the phrases "significant increase" and "significant decrease." Similar to 1990, however, was the use of confusing phrases such as "1% to 37% were Proficient." It was not clear if this meant just Proficient, or at or above Proficient.

Additional variables highlighted in the 1992 report included race, region (which was not mentioned in 1990), sex, and private school versus public school student comparisons (also not mentioned in 1990). Variables highlighted in 1990 but ignored in 1992 included the type of community (SES), parental education, and number of mathematics courses (grade 12).

The statistic highlighted most among speakers at the press conference for the 1992 Mathematics results was the finding that 8 of the 37 States had significantly improved their average scaled score from 1990 to 1992. Furthermore, no State had a significant decrease. A popular quote in 1993 that appeared in many newspapers around the country was, "The Proficient level is the one we are really shooting for" (Mary R. Blanton, a member of the NAGB board). Overall, the mathematics scores were all better than in 1990.

1996 Mathematics. The main focus of the 1996 Mathematics press release materials was the improvement in the average scores across all grades over time (since 1990). An important innovation of this press release was the increased focus on information concerning NAEP itself and the mathematics framework. The content areas measured by NAEP as well as the achievement levels were defined. Reporting of this type of information seemed to be an important strategy that could help members of the press better understand the goals and purposes of NAEP and NAGB. In turn, this effort held the potential that the amount of misinterpretation of NAEP results by the press could be reduced. Also new in 1996 was an increase in the use of tables and figures. Unfortunately, those graphics were not available for this report.

Unlike in 1990 and 1992, average scaled scores were reported for the Nation and for each grade. In addition, scaled scores were reported for grade 4 from 1990 to 1992 to 1996. This switch in reporting was somewhat surprising, although much more emphasis was placed on the achievement levels throughout the data presentation. The percentage of students in each grade at or above Basic was presented from 1990 to 1992 to 1996. Improvements ranged from 10 to 15% (in grade 8 and grade 4, respectively). Only grade 4 showed significant improvements at the Proficient level, the percentage going from 18 to 21%.

Surprisingly, only race and sex variables were highlighted in this report. There were some changes in the performance of boys and girls but no change in the patterns related to race. There was no difference in average scaled scores between boys and girls in 8th and 12th grades; however, 4th-grade boys had a slightly higher average than 4th-grade girls. There was a gap between 4th- and 12th-grade boys and girls at or above the Proficient level. Whites were still performing below the levels of Asians and above Hispanics and African-Americans. The gap in performance was still as large in 1996 as it was in 1992 and 1990.

State results were also reported. In grade 4, 15 of 39 States increased in average performance, 3 States declined in performance, and 21 had no change in average performance. For grade 8, only 13 of 37 increased, none declined, and 24 had no change in average performance. The good news at the time was that 27 of 32 jurisdictions improved since 1990.

Information related to regions, type of community (SES), parental education, number of mathematics courses taken, and comparative results for public versus private schools was missing. We were surprised that none of these variables were highlighted in the data presentation. The presentation did seem more narrow in focus compared to 1992 and, as a result, easier to digest.

Statements by others, including Secretary of Education Richard Riley, seemed to focus on the positive trends in both the national and State results. He singled out North Carolina, Michigan, and Texas as most improved at the grade 8 level. He cautioned, however, that all the news was not good. Secretary Riley stressed how poorly U.S. children looked compared with other nations based on the Third International Mathematics and Science Study (TIMSS) results. He mentioned that 8th graders were below average internationally and that compared with other countries our mathematics curriculum was less challenging. A quote that appeared at the time was "give every child a world-class education." International comparative information had entered NAEP reporting.

Analyses of News Clippings for Mathematics: 1992 and 1996

1992 Mathematics. Ten articles reporting the 1992 Mathematics NAEP results were carefully analyzed. Five of the articles were written for large or nationally oriented newspapers and five appeared in smaller local papers. Before comparing and contrasting the contents of these articles, a brief review of the information released by NAGB on April 8, 1993, is provided.

The basic message released by NAGB during the press conference was that mathematics performance was increasing; however, only a small percentage of students were able to do complex problem solving. Topics emphasized in the press release materials included national and State performance, change in achievement from 1990 to 1992 for the Nation and States, the achievement levels (which were new), and actual student test items as examples of what students should be able to do. In addition, performance was presented in terms of race/ethnicity, region, sex, and public versus private schools. The report did not mention how type of community, parental education, and number of mathematics classes related to performance. At no time in the

press conference were average scaled scores presented, because the emphasis was on reporting scores in terms of achievement levels.

Out of the 10 articles, only one (the *Washington Post* article) focused exclusively on national results, and no reference was made to State results or other variables, including race and sex. Half the articles, mostly in local newspapers, focused mainly on their own State's performance. These five articles provided either no coverage of national results or coverage of national results solely for comparison to the State. Three articles had a balanced focus on both national and State results. Finally, one article (in the *Wall Street Journal*) was not about the test results. Instead, the article explored the question "Do mathematics games help students learn mathematics?" In this case, the 1992 NAEP mathematics results were used as data to support the author's thesis.

Nine of the ten articles used achievement levels to describe student performance. The one article that did not use the words Basic, Proficient and Advanced (the *Philadelphia Inquirer*), described the students' performance in terms of average scores (e.g., "increased from 262 to 266"). Six articles used both achievement levels and scaled scores. The most common use of scaled scores was to show changes in average score from 1990 to 1992. Three articles used achievement levels exclusively to describe student performance. Only 2 of the 10 articles defined all three achievement levels, although, one article defined the Proficient level by quoting Mary Blanton, one of NAGB's board members.

Half the articles mentioned differences among racial groups. The *Boston Globe* article was the only one that dedicated nearly half of its column space to the performance of minorities and students from poor communities. Only three articles mentioned differences in performance based on sex. Only one article mentioned differences in performance by type of community. No articles mentioned the relationship of parents' education level and scores on the 1992 NAEP Mathematics Assessment. Seven articles tied the increase in performance to changes in the curriculum. Several authors quoted Secretary Riley's statement that the report provided "early evidence that challenging curriculum, standards, and assessments can work to improve student performance."

A popular explanation in the articles for why students were doing better or worse in mathematics was television. The newspapers picked up on the relationship between watching television and performance on the NAEP Mathematics Assessment. Four articles discussed the negative effect of television on mathematics proficiency. Obviously, the newspapers were unaware that a high correlation does not establish a causal relationship. In States where improvement occurred, such as North Carolina and Colorado, it was stated that this was because students were watching less television. In one of the Florida newspapers, much of the blame was placed on television while ignoring all other variables related to achievement, including SES, parents' education level, the curriculum, and more. Selective reporting of results, perhaps to support writers' favorite explanations for the findings, was common. Interestingly, none of the newspapers drew attention to the fact that increases in NAEP Mathematics results were positively correlated with the price of newspapers!

Generally, the numbers reported in the articles appeared accurate. One exception was the report that "over a 2-year period, 34 of 37 States showed increased proficiency in 8th grade." The statement should have said that 18 of 37 States made statistically significant increases in performance. In one or two cases, the word "substantially" was substituted for the word "significantly"; however, a substantial increase is not the same as a statistically significant increase.

The problem of misrepresentation was not related to the numbers. There was a problem with the language used to report achievement levels. To describe students that did not reach the Basic level, articles used the following phrases: "failed to exhibit basic competence," "were not able to solve basic mathematics problems," and, "4 out of 10 can't handle basic mathematics." These phrases seem to be technically inaccurate, invalid, and somewhat misleading. It seems that several of the writers did not completely understand what these levels of achievement represented in terms of mathematics skills. Only one or two articles used the example items to describe what a Basic or Proficient student could do. One writer changed the definition of Advanced to "above average." So, although there was substantial evidence that the media were using the achievement levels to report performance, the meaning of the achievement levels was not being reported clearly. The consequence was substantial misinformation on this point in the newspaper reports of NAEP results.

1996 Mathematics. Ten articles reporting the 1996 Mathematics NAEP results were analyzed. Four of the articles were written for medium to large national newspapers, and six appeared in smaller, local papers. Before comparing and contrasting the contents of these articles, a brief review of the information released by NAGB on February 28, 1997 is presented.

The main focus of the 1996 Mathematics press release materials was the improvement in the average scores across all grades since 1990. There was increased focus on information concerning NAEP itself and the mathematics framework. We expected this type of information to help members of the press better understand the purpose of NAEP and to reduce the amount of misinterpretation of NAEP results. Unlike 1990 and 1992, average scaled scores were reported for the Nation for each grade. Despite examples of reporting using scaled scores, results were primarily described in terms of achievement levels. Only two variables, race and sex, were highlighted in this report. Information related to regions, type of community (SES), parental education, number of mathematics courses taken, and public versus private school students was missing. Comments by Secretary Riley stressed how poorly U.S. children had performed compared with children from other nations, based on the TIMSS results.

Newspaper reporting trends were fairly straightforward in the 1996 results. The main message printed by the large papers was identical: "Scores are improving nationally, but we still need to do better." Three of the four large newspapers, the *New York Times*, *USA Today*, and the *Philadelphia Enquirer*, focused primarily on national results. Only the *Washington Post* had a balanced focus on national and State results, reporting the performance of students in Maryland and the District of Columbia. Six smaller newspapers focused primarily on their own States' performances and provided either no coverage of national results or coverage of national results solely for comparison to the State. For each State, it was either good news (improvement) or bad

news (below the national average or at the bottom). This polarization of focus was a subtle change from reporting trends in 1992.

The most noticeable change in reporting Mathematics trends from 1992 to 1996 was the decrease in the number of articles using achievement levels to describe student performance. In 1992, 9 of 10 articles used achievement levels. In 1996, only 6 of 10 articles used the achievement levels. This may be due to the increased emphasis on scaled scores rather than achievement levels in the 1996 NAEP press release materials. In addition, none of the articles defined the achievement levels, and only two of the articles discussed the achievement levels beyond simply using the terms Basic, Proficient, or Advanced. Three articles used scaled scores exclusively to communicate student performance. Finally, three articles used both achievement levels and scaled scores to report student performance. Again, the most common use of scaled scores was to show changes in average score from 1990 to 1992 to 1996. As in 1992, only three articles used achievement levels exclusively to describe student performance.

Only sex and race variables were presented at the press conference in 1996. Surprisingly, only two papers (the *New York Times* and the *Washington Post*) mentioned differences among racial groups and between boys and girls. Both articles reported the relationship of both variables to student performance. In 1992, half the articles mentioned differences among racial groups. The lack of interest by the press in these and other variables related to student achievement seemed to be declining. Only one article addressed the relationship between poverty and test scores, and only one article mentioned the relationship between parental education and test scores. Additionally, only one article highlighted the similar performances of public and private schools. Eight of 10 articles emphasized the need to improve the curriculum and half mentioned the performance of students on NAEP and TIMSS. Essentially, performances on these tests led the press to conclude that students were doing better but not performing at a world-class level.

Teacher training was mentioned in conjunction with improved curricula as the answer to mediocre test scores. The consistent cry of the local politicians was, “we need to continue to do a better job training our teachers and creating a more challenging curriculum in order to improve our test scores.” No articles discussed the limitations of NAEP, and only one mentioned other subjects measured by NAEP. In general, there was less information about NAEP and NAGB compared with previous years.

For the most part, it appeared that the numbers reported in the articles were accurate. Although to a lesser extent than in 1992, there were still problems with the language used to report achievement levels. One example was in a Tampa, Florida, newspaper that read, “nationwide, 20% of 4th graders passed.” This was a gross misrepresentation of the results; the Proficient achievement level was not considered passing. In the *Philadelphia Enquirer,* an example was found in which the percentage of students at different achievement levels was reported, but the author avoided using the actual words Basic, Proficient, and Advanced. And, finally, some newspapers were still describing students who did not reach the Basic level as people who, “lack basic mathematics skills” or “lack skills in mathematics.” Although these are not examples of gross misinterpretation, they do highlight representations of the results that are technically

inaccurate, invalid, and somewhat misleading. It appears that the authors of these articles did not completely understand the meaning of the achievement levels.

Only 3 of 10 articles used graphics in their stories. The North Carolina paper presented a table of the percentage of 4th- and 8th-grade students at or above each achievement category nationally in 1992 and 1996; however, the graphic did not tie in directly with the text. The *New York Times* used a bar chart for each grade (4, 8, and 12) to show the percentage of students testing at or above Basic for 1990, 1992, and 1996. Finally, *USA Today* printed a table containing the percentage of students at or above Basic and below Basic for all participating States in 1992 and 1996, however, the article didn't indicate the grade represented in the table!

Comparison of Press Releases for Reading: 1994 and 1998

1994 Reading. The 1994 Reading press release offered something new. In addition to defining the NAEP score scale and the achievement levels, Commissioner Emerson Elliott defined the term "statistical significance." How the press understood the term and used it in their stories was interesting. The foreseeable problem with using the term is that when it is not used to describe change (decline in scores), it is not always clear whether the change was statistically significant. This could lead to more frequent misinterpretations than if the term had not been introduced.

The 1992 reading results (where achievement levels were not used in score reporting) were disappointing. The 1994 reading results were also negative in all groups and all grades. National results were presented in two separate sections, one based on average reading proficiency (scaled scores) and the other based on achievement levels (standards). Nationally, there were no changes in average proficiency for 4th- and 8th-grade students; however, there were statistically significant declines for 12th-grade students. In addition, 10 of 38 States or jurisdictions declined significantly. It was not clear, however, that the press was distinguishing between increases and decreases and significant increases and decreases. Four variables were discussed: sex, race/ethnicity, parental education level, and public versus nonpublic school student performance. There were no changes in patterns from 1992 to 1994 for these variables. There was a somewhat awkward attempt at explaining what the large gaps in performance among racial/ethnic groups meant. It was suggested that average overall difference between 4th and 8th grades was 45 points and between 8th and 12th grades was 26 points. Because no scaled scores were reported by race/ethnic group, it was not clear how average gains between grades related to gaps among the race/ethnic groups. It is unclear whether the goal was to gloss over the findings on race/ethnic groups. State results were also mentioned. The 10 States that performed best were named.

Findings based on achievement levels also indicated declines in student performance. Average performance in 27 States was Basic and in 14 States was below Basic. There were only two significant improvements among States: (1) the percentage of Advanced students increased in Arizona, and (2) the percentage of students at or above Proficient increased in Mississippi. Additional information was presented, assuring the media that the shocking decline in students' reading ability was real and not a methodological artifact.

Another speaker (James Ellingson) also stated that reading proficiency was poor and that no reasons were apparent from the NAEP assessment. At the same time, he argued that something must be done about the dismal performance. The numbers were emphasized in the report: 10 States showed significant declines and none improved. It was mentioned that parents needed to promote reading in the home, and less television should be watched. At school, it was believed that students should have access to literature and must be encouraged to read. When speaking about the NAEP results, William Randall, a member of the NAGB board used the catchy phrase, the results are a wake-up call, or a "whack on the head." Secretary Riley said, "Too many students are spending too little time reading and too much time watching mind numbing television." Many media outlets picked up these quotes. Presumably, to have a result quoted, it should be colorful (e.g., "Mathematics results add up to failure").

1998 Reading. Compared with the 1994 results (which were relatively brief and efficient), the 1998 Reading press release materials were exhaustive. In fact, there may have been too much information to process. In addition to the usual data presentation and statements by key speakers, several additional documents were distributed, including a handout containing sample questions, an Executive Summary report, and a 12-page newspaper report. These were only the national results. It was not possible to review all the material for this report. Only the data presentation and comments by key speakers are considered here. The central message of the 1998 report was simple (despite the complexity of the materials): there were some improvements in grade 8 (since 1994) but, overall, not much progress was made.

Commissioner Pat Forgione's data presentation began by describing the three types of reading being tested: (1) reading for literacy experience, (2) reading to gain information, and (3) reading to perform a task. Unfortunately, he did not define these three reading types or describe how they were used in the reading assessment. This seemingly limited the usefulness of the information. Also new for 1998 was the organization of the results. The results were presented separately for each grade (4, 8, and 12). Within each grade, results were broken down in terms of scaled scores and achievement levels. Within each grade, results were presented by the following subgroups: sex, race, and public versus nonpublic schools. Unlike 1992, no mention was made of the relationship between parents' education level and student achievement. For grades 4 and 8, changes in scaled scores (+4 or +8) were used to describe changes from 1994 to 1998. Actual scaled scores were used only in grade 12 (292 to 287 to 291). It seemed strange to describe changes in scaled scores without actually giving the scores themselves.

On top of this information, Commissioner Forgione described data related to four other variables: (1) daily reading habits, (2) reading and writing for school work, (3) explaining understanding and interpretations, and (4) television viewing. A separate executive summary report presented information on additional variables, including parents' education level, region of the country, type of location, and free/reduced-price lunch program. The report also described several school and home factors (in addition to those outlined by Commissioner Forgione) that contributed to reading performance, including reading self-selected books in school, discussing studies at home, and talking about reading with family or friends.

This tremendous amount of material made available to the media probably provided them with more information than they could use in any one article. There seemed to be no attempt to focus them on any particular findings. One wonders about the desirability of providing the press with substantial amounts of data with little guidance on their meaning. At the same time, highlighting many variables and factors might lead the press to appreciate the complexity of the issues related to reading education. Unlike the more positive Mathematics reports, where "too much television" was viewed as the main excuse for poor performance (blamed for all of America's education problems), hopefully, the media would be less likely to draw inappropriate conclusions about the causes of poor reading test scores.

Analyses of News Clippings for Reading: 1994 and 1998

1994 Reading. Ten articles reporting the 1994 Reading NAEP results were selected for close analysis. Five articles were written for large or medium nationally oriented newspapers, and five represented smaller local papers. A review of the 1994 NAEP Reading results presented by NAGB on April 28, 1995 is presented first.

The message consistently voiced by the large newspapers was that, nationally, scores for seniors were dropping significantly. Two of the five large papers mentioned "there is no explanation"; all the large papers agreed that something needed to be done. Only one newspaper, *USA Today*, focused exclusively on national results. As with mathematics reporting, the smaller newspapers tended to focus on their own State's results.

In 1994, newspapers tended to report NAEP results in terms of achievement levels: 7 articles used the achievement levels, and 3 did not. This number was less than the 9 out of 10 articles using achievement levels for the 1992 Mathematics test results. Also, in 1994, only 3 articles used scaled scores, and 7 did not. The fact that the press release described performance in terms of average reading proficiency made this finding unusual.

The most popular variables mentioned in the newspaper stories included race (7 of 10), private school versus public school (6 of 10), and sex (5 of 10). Although highlighted in the press release, no article mentioned parents' education level. None of the articles mentioned the relationship between poverty and test scores or any other test scores to NAEP. Only 3 of 10 articles made reference to the other subject areas (mathematics and science) included in NAEP. No stories discussed the limitations of NAEP (e.g., no student or school scores). Surprisingly, few stories used the information on other factors related to reading.

For all subjects and all years, the newspapers have generally done a good job of presenting the numbers. They seem able to communicate accurately the percentage of students that are below Basic and the percentage that reached Proficient. They are also equally adept at listing average scores and indicating the number of points gained or lost over time. This is probably a compliment to those preparing the press release material. The big problem is that the press is not adequately communicating the meaning of these numbers. In some cases, it seems that NAGB and NCES could do a better job of briefing the press, and, in some cases, it is the press' own desire to dispense credit and place blame that leads to misinformation.

First, we focus on the tendency of the press to draw erroneous conclusions. Because nearly all of the 1994 Reading results were poor, the newspapers tended to point fingers or make unsubstantiated causal inferences. For example, *USA Today* was guilty of making an unfounded connection between race/minorities and English as a Second Language (ESL) students. In one sentence, the paper reported that 40% of whites, 30% of Asians, 18% of Hispanics, and only 12% of African-Americans reached the Proficient level. In the next sentence, they said, "Tests don't accommodate students whose home language is not English." In no way can such a statement (that some students taking the test do not speak English as their first language) explain or shed light on the complex factors that have led to differences among racial groups in this country. The implications are that NAEP is not valid and that most minorities cannot speak English. One other article (*Richmond-Times Dispatch*) directly called into question the validity of the test. The writer quoted the State division chief for testing by writing, "Given the random nature of the test, she and others are trying to assess whether it truly indicates students' performance [in reading]. 'The truth is we don't know the answer,' she says." By making that statement, the whole State could discount the poor performance of students in the State.

Some additional factors related to reading proficiency were outlined in the press release (students are reading less for homework), but hardly any articles utilized this kind of information. Instead, the media preferred to highlight the relationship between poor reading and too much television. This common problem among the articles may have been partially the fault of Secretary Riley, who spun a catchy quote that too many kids are spending too little time reading and too much time watching television. Popular in the more politically oriented articles was a focus on (1) the amount of money spent on education, (2) the poor preparation of teachers, and (3) problems in the curricula. Finally, the media appeared to be easing up a bit on schools and starting to place blame on parents in the home.

By improving the quality of its press release material, NCES and NAGB can address the way the press understands test scores. The media appear to be having some trouble with the "statistically significant" difference phrase, and they struggle with the scaled scores and the meaning of a 3-point change versus a 5- or 8-point change. How large a change is significant or meaningful? Commissioner Elliott discussed the gap in performance among racial groups by indicating that the difference between 4th and 8th graders is 45 points and the difference between 8th and 12th graders is 26 points. A few more sentences could have tied that information explicitly to the differences among racial/ethnic groups, and then the public would have realized how serious the differences are among various subgroups. Unfortunately, the relevant connections were not made, so the correct interpretation was never made clear.

Currently, news articles report facts such as an 8% improvement and a 3-point gain. Unfortunately, it seems that no one (the writers and the readers) is sure whether these are significant, or what the term "statistically significant" means. *USA Today* reported that scores for all grades declined since 1992. In reality, only the scores for 12th graders showed statistically significant declines. Many newspapers can handle this technical term; however, NAGB should continue to try to clarify the meaning of these concepts for the press.

1998 Reading. Ten articles reporting the 1998 Reading NAEP results were analyzed. Only 2 of the articles were written on February 11, 1999, and they focused primarily on national results. The other 8 articles were written by small papers and focused on State results. The message voiced by most papers was that, nationally, scores were up slightly, but more needed to be done. Small papers tended to focus on gains or losses by their own State.

In 1998, most newspapers tended to report NAEP results in terms of achievement levels: 7 articles used the achievement levels, and 3 did not. This number remains the same as in 1994. Five articles, however, used scaled scores, an increase of 2 from 1994. It seems this increase in the use of scaled scores reflects the increased prominence of scaled scores in the press release. Also, seemingly a result of the way scores are reported in the press release, more newspapers reported gains or losses using plus or minus the difference (e.g., "the average grade 4 reading score was up from 1994 (+3 points)"). Whether a 3-point change is statistically significant was not made clear.

Overall, few variables were mentioned in the articles. Only 3 mentioned race, only 3 mentioned sex, one mentioned socioeconomic status, and none mentioned private school versus public school student performance. The choice by newspapers to ignore these variables may have been related to the 1998 State results, which only contained a few sentences about race/ethnicity, sex, and participation in a free school lunch program. In other words, these variables were down-played in the State results compared with national results (or possibly these results were to be released at a later date). Only 2 of 10 articles made reference to the other subject areas (mathematics and science) tested by NAEP. No articles discussed the limitations of NAEP. Few articles used the information on other factors related to reading. Again, this is probably because of the content of the 1998 State press release materials.

The criticisms of the findings in the 1998 news reports were similar to those presented above (with the 1994 newspaper analyses): Statistical terms and the meaning of scores and score differences were not made clear. Often newspaper stories would make the comment that NAEP scores ranged from 0 to 500, but this tells readers very little. It does not provide adequate background information to give meaning to the scores. Suppose a reader knows nothing about NAEP and reads that the average score in Florida for a 4th grader is 207, which is down 1 point from 1994. Also, the Florida State average of 207 is 8 points below the national average (215). Then the reader learns that the average score for 4th graders in Connecticut is 232. Did Florida really decline, since 1 point may just be due to chance? Should the reader be troubled by the fact that Florida is 8 points below the national average? Should the reader be more concerned that the difference between Florida and Connecticut appears large (25 points)? What does an 8-point difference or a 25-point difference really mean? How large a gain or loss is statistically or practically significant? These are the types of questions that appear worthy of attention in future releases of NAEP results.

Even a person with some statistical knowledge would have difficulty making sense of the results because score distributions and basic descriptive statistics are not given. Adding to the difficulties are the unequal differences in gains between grades (45 points between 4th and 8th

and 26 points between 8th and 12th). Complicating matters further are the difficulties in correctly interpreting gain scores and the recent switch in scale (from 0 to 500 across grades to 0 to 300 within a grade). The same problem exists when talking about percentage gains or losses. Connecticut schools increased the percentage of children who reached Proficient by 8% since 1994 and by 12% since 1998. Are these gains significant, and in what sense?

Press Release for Science: 1996

The 1996 Science press release materials were different from most previous NAEP press release materials in that they reported results exclusively in terms of achievement levels. (Presumably this was because of the controversy surrounding the 1996 Science achievement levels. The consequence was separate releases of the NAEP scaled scores by NCES and NAGB.) Dr. Mary Lyn Bourque, the measurement specialist on the NAGB staff, did mention that the scale went from 0 to 300, but that was the only time scaled scores were mentioned. Some background was provided about the nature of the NAEP assessment. It was described as hands-on in nature and consisting primarily of open response-type items. In addition, she noted that three areas of science were measured by the assessment: physical, life, and Earth. Then, in a straightforward manner, the test results were presented, first for the Nation and then for the States. No comparisons were made to previous science assessments. Results (nationally and by State) were broken out by sex, race/ethnicity, parents' education level, and type of school (public versus private). The presentation of results was clear and concise.

Secretary Riley's remarks may have been less clear. First, he mentioned that 4th-grade students in the United States performed second in the world only to Korea in science on the TIMSS assessment. Then he stated that only 67% of the U.S. 4th-grade students scored at the Basic level or above. These two findings seem to be contradictory. One plausible interpretation of his remarks would be that achievement levels at grade 4 were set too high in that the TIMSS assessment results seemed to suggest that U.S. students were, on the average, already world class. How would the Korean 4th graders do on the NAEP Science Assessment? Would only 72% of their students reach Basic? However, there was no mention of potential problems with the achievement levels in the newspaper reports.

Secretary Riley went on to say that the TIMSS results indicated that U.S. 8th-grade students did poorly compared with other nations in chemistry and physics. The question of whether U.S. science curricula cover chemistry and physics in elementary school was not addressed. Mark Musick, a NAGB board member, discussed the 1996 NAEP Science achievement levels. He appeared to have made a good statement in defense of using achievement levels, when he said (in regard to scaled scores), that half the students score below average, and the average (sometimes) can be woefully inadequate.

Analyses of News Clippings for Science: 1996

Five articles reporting the 1996 NAEP Science results were carefully reviewed. All five of the articles were written for large or nationally oriented newspapers. The unique features of the press release were that no results were reported in terms of scaled scores, and no previous science

results were available for comparison. In addition, TIMSS results were released shortly before the 1996 NAEP Science results, so there was a basis for interpreting the NAEP Science results and the TIMSS results together.

Of five articles, four focused on both national and State results. Only one article focused exclusively on national results. There were no comparisons to previous science results (none were released). TIMSS was mentioned in only one or two of the articles. The main point in each story was that American students were not competent in science, despite the performance of 4th-grade students internationally.

All five articles reported scores in terms of achievement levels to describe student performance, although only two newspapers tried to explain the meaning of the performance levels. For the first time, none of the articles used scaled scores. This is probably a direct result of the way data were presented at the press conference. Another explanation might be the fact that, in other press releases, the most common use of scaled scores was to show changes in average scores over time (e.g., from 1994 to 1998).

Three articles mentioned other variables; two did not. Differences among racial groups were mentioned in two articles. Two articles mentioned differences in performance based on sex, and two mentioned differences related to parents' education level. Only one article mentioned differences in performance by type of community (or SES).

Three of the 5 articles contained figures. The figure in the *Washington Post* was a graphic item that asks, "What ripple pattern will form if a stone is thrown into a pond?" The figure had the following caption: "Two in five fourth graders didn't know the answer was C." The other two figures were similar in content, but slightly different in organization. Both *USA Today* and the *New York Times* printed tables containing the State results for 8th graders. Both tables were in alphabetical order, although *USA Today* put the national average at the top of the column, and the *Times* put the national results last. The columns (from left to right) in the *USA Today* table are State, Average Score, Advanced, at/Above Proficient, at/Above Basic, and Below Basic. The columns (from left to right) for the *Times* are: Less Than Basic, Basic, Proficient, and Advanced. A line was used to group together the columns that are Basic or better. The *New York Times* presentation appeared clearer. The well-known problem with cumulants existed with both tables (Wainer, Hambleton, and Meara, 1999). Ordering results alphabetically by States could also be criticized.

There seemed to be fewer inappropriate conclusions drawn by the press in these articles than in some of the previous releases, possibly because larger newspapers were less likely to make erroneous statements than smaller papers (which were not represented in our sample).

Press Release for Writing: 1998

The 1998 Writing press release material was very informative. Dr. Gary Phillips, the Acting Commissioner of NCES, provided a direct and unequivocal statement at the beginning of his remarks: "The average or typical American student is not a proficient writer." Correlational

evidence with writing scores was also presented in a meaningful way. For example, it was noted that students who provided visual evidence of planning their work tended to score higher on the writing assessment. Results by region of the country were highlighted in the press release. Similarities in the findings to other NAEP releases were also part of the presentation. Breakouts of results for all the major demographic variables used in the past were used again in reporting the writing scores. At the 8th-grade level, States were compared in writing proficiency. Lots of valuable information was presented using very simple graphics.

One of the major findings in the study was the results for boys and girls. By State and by grade, girls outperformed boys. These sex findings were then compared with sex results in other subjects. This comparison seemed especially helpful. For example, knowing that the differences between boys and girls in writing skills exceeded any differences obtained in other subjects seemed significant. The press release material went on to highlight the correlations of writing scores with approach to writing instruction, planning for writing, and school and home factors. Finally, suggestions for parents to enhance the writing skills of their children were offered. One speaker at the press conference (Marilyn Whirry, a NAGB board member) spent time addressing behaviors of teachers that might improve writing proficiency. Another speaker (Richard Sterling, Executive Director, National Writing Project) realized the importance of instruction in developing writing skills and said much was known from research about developing writing skills. He also emphasized the importance of the integration of writing skills with other subjects. The NAEP 1998 Writing Report Card Highlights was especially well done. The language was clear, the graphics were excellent, and the interpretations of the findings were straightforward and informative. In almost every way, this document appeared to be more informative than the corresponding document that had been prepared for 1996 NAEP Science, which was the first of this type of report aimed at the "person on the street" (see Hambleton and Smith, 2000).

Analyses of News Clippings for Writing: 1998

The newspaper reports we studied carried similar stories—the results were not very good, and ways must be found to improve writing skills. A typical comment from the newspaper stories was, "Last week's report that three-quarters of American schoolchildren failed to score at the Proficient level in a national writing test has raised serious questions about how to improve their skills." Typical in the newspaper stories were references to the achievement levels and their descriptions and to the sample questions that were released. Reporting the sample questions and some sample answers was more prevalent in this release than others we had looked at. Again, it was common to highlight the demographic breakouts of results—by sex and race/ethnicity, and so on, but some attention seemed to be focused on approaches for increasing writing skills. For some, the approach must include portfolios, regular conversations with teachers about writing, writing drafts, and organization of thoughts before writing. For others, the solution was to "stick to basics" and focus on writing skills in the early grades. These were reasonable suggestions given the reported data.

State-to-State comparisons and State-to-Nation comparisons were popular topics for the newspaper stories. The headline in one Austin, Texas, paper was "Texas eighth graders

outperform the Nation," with comments from the Texas Commissioner of Education on the significance of the high ranking. (Actually, Texas did well in the State rankings, but it was not first; on the other hand, Texas did exceed the national average, but so did many other States. Of course, it would be incorrect also to imply that all Texas students exceeded the national average.) In a typically performing State such as Tennessee, the headline was, "Tennessee writing scores keep pace with national average," whereas in a low-performing State, the lead was, "Stick to basics," and the excuses (for low writing scores) "stink." The overall reporting of the writing scores was accurate and informative. Presumably the excellent press release materials were a major factor in the quality of reporting.

Summary of Main Findings and Conclusions

Our findings revealed a complex pattern of NAEP score reporting by the press since 1990 but, in general, score reporting appears to have improved considerably over the time period. The complexity is due to three factors: (1) single versus joint releases of NAEP results, (2) national and State NAEP reporting, and (3) unique features about several of the NAEP releases. First, several of the releases do not appear to have been jointly sponsored by both NCES and NAGB and this may explain the fluctuations over time in the emphasis given to scaled scores (e.g., 1996 Science). Second, some of the NAEP releases provided both national and State results and others provided only national results. This feature in the NAEP design made it more difficult to follow reporting trends. Finally, there was something special or unique for several NAEP releases. For example, the 1990 Mathematics was the first use of achievement levels; the 1992 Mathematics provided the first opportunity to assess gains with achievement levels; the 1996 Science and Mathematics results could be compared with results from the 1995 TIMSS; and the 1998 Writing was the first use of achievement levels in Writing and the first time students tested with accommodations were included in the overall results. These unique features were often highlighted in the NAEP press briefing materials. Therefore, in attempting to answer the four main questions of the study below, general impressions are substituted for precise counts of observations that were made in reviewing the press release materials and the newspaper stories.

The study was designed to answer four questions. Our answers follow, along with a set of conclusions.

1. How have the NAEP press briefing packages changed over the past 10 years?

Using the Mathematics reporting in 1990, 1992, and 1996 as our major database to address the question, it appeared that the press briefing packages were changed substantially over the three releases. Over time, more information was given about NAEP itself and the curriculum frameworks and content. In addition, exemplar items were introduced, along with more use of graphics in score reporting. We noticed similar trends in the 1996 Science and the 1998 Reading and Writing. In later NAEP releases (1996 and 1998), the press briefing packages appeared to be more informative—NAEP frameworks, content information, and sample items were made available. Also, more figures, graphs, and tables appeared to be used. The newspaper-like reports with the 1996 Science, 1998 Reading, and 1998 Writing appeared to be major improvements in reporting results over more technical reports (called "Executive Summaries") that were available

with earlier releases. Finally, the press releases generally appeared to be more focused in later releases (1996 and 1998), and the points highlighted were done so in depth. That is, there appeared to be fewer demographic breakouts of the data, but those provided were discussed in considerable detail. The principle that less might be more seemed to be operative, so breakouts of the data by region, for example, were less common. We also believed, but did not take the time to check fully, that, with the 1998 Reading release, a serious effort was made to interpret the findings and suggest directions or remedies for the improvement of reading skills. This appears to be a very positive change in NAEP score reporting.

Clearly, the debate about the validity of the achievement levels is not a central issue in newspaper stories today. After newspaper stories in 1990 and 1992, there were few references to the debate, and few questions were being raised in the press about the need for additional validity evidence to support the achievement levels. Of course, the topic of NAEP achievement level validity continues to be important for policymakers and researchers (see, e.g., Hambleton et al., 2000; Pellegrino, Jones, and Mitchell, 1998).

2. What information has been highlighted in the newspaper accounts of NAEP results?

Not surprisingly, the press release materials are the main source of information for newspapers. Literally everything that is included in the press release material was found in someone's newspaper story. They report the information they are given. Also, newspapers appear to focus on State rankings when they are available; achievement level reporting; and sex, ethnic, and private school versus public school student breakouts of data. Barron and Koretz (1998) reported a similar finding from their research. These items were frequently included in the newspaper stories we reviewed. Finally, the smaller newspapers appear less interested in national results and are more likely to focus on information they think will be of greater interest, such as State results or possible interpretations of the results (e.g., the role of television in school achievement).

3. Is there evidence that the NAEP press release materials are being understood and used by the newspapers in their stories?

There is substantial evidence to suggest the NAEP press release materials are being used and understood by the press. The problems occur when the press tries to go beyond the materials provided and interpret the findings for the public. A careful reading of the NAEP releases leads us to feel that NCES and NAGB may not want to interpret the complex array of findings for the public; they may want readers to arrive at their own interpretations. However, this goal may be unrealistic because (1) quantitative literacy is not high across the country (Kirsch et al., 1993), and (2) the data and NAEP reports themselves are very complex. One idea might be to model correct interpretations from the data. For example when Secretary Riley connected the improvements in Mathematics performance in 1992 over 1990 to changes in the mathematics curricula, he might have offered other explanations that would be consistent with the available evidence and pointed out the difficulties in making causal inferences from correlational data. Suggestions for followup research might be offered as ways to definitively establish the correct explanation or explanations of the findings.

Also, some improvements in the communication of results possibly could be made by NCES and NAGB. For example, the achievement level descriptions may still be too cumbersome and vague to be fully understood by the press and the public. Simpler descriptions, with focus on the differences in student knowledge and skills at each proficiency level, and with more sample items, may be worth investigating as a solution to the problem. At the same time, substantial improvements in communicating the meaning of the achievement level descriptions appear to have been made since 1990.

4. Are the newspapers accurately conveying information about NAEP results to their readers?

The meaning of achievement levels continues to be misinterpreted by the press, and statistical misinterpretations remain, such as inferring causality from correlational information. Statistical significance has been interpreted as substantial significance, but this may be untrue in many instances. In 1992, a negative correlation was reported between the amount of television watching and mathematics results. The press immediately wrote about the need for low-performing students to turn off their television sets. However, watching television may be a proxy for many other explanations of low mathematics scores, including lack of student motivation, student interests in other activities besides mathematics, and lack of parental supervision. Turning off the television set may have no effect at all in raising mathematics scores. It seems that the newspapers are going to try to explain NAEP results. It might be better if NCES and NAGB offered explanations whenever possible, highlighting both correct and incorrect interpretations, and describing research that might help in resolving questions about correct interpretations.

Our research findings lead to three conclusions about NAEP score reporting by the press. First, our impression is that the press release materials over the past decade have improved considerably. The press release material for NAEP Writing in 1998 was considerably better than NAEP Mathematics in 1990. In 1998, there was more use of graphics; information about NAEP and its goals; and information about the frameworks, curricula, and exemplar items. All of these features are likely to contribute to the understandability of NAEP score reporting. At the same time, there seemed to be fewer demographic breakouts of data in the 1998 reports than in the 1990 reports; therefore, the press release materials appear to be more focused on presenting a simpler and clearer message of the findings. (1998 Reading may have been an exception to the trend that we observed.) More standardized press release material from one NAEP release to the next may also contribute to understanding and effective communication of NAEP scores because there would be fewer novel data presentation formats for the press to handle. However, we understand and appreciate the tension that surely exists between being consistent in reporting NAEP results over time and capitalizing on innovations in score reporting.

Second, the newspapers appear to be able to report the numbers provided by NCES and NAGB correctly, but some problems in NAEP score reporting were evident. The achievement levels have generated some interest in NAEP scores but still do not appear to be fully understood. “Above Average” is substituted for “Advanced”; Basic students have been described as “basically competent.” Language and examples need to be found to communicate the correct

interpretations of Advanced, Proficient, Basic, and below Basic. What are the knowledge and skills possessed by students at each level, and what are the differences among the performance categories? These appear to be two of the questions that need to be satisfactorily answered to improve the reporting of NAEP scores.

Finally, problems remain in NCES-NAGB efforts to explain the meaning of statistical concepts and scores. Much of the statistical jargon that was associated with NAEP reports before 1994 is gone. Still, terms such as "statistical significance" have been used in more recent reports and do not appear to be understood by the press. Percentiles and cumulative percentages are two more statistical concepts with a history of being misinterpreted by the press. The confusion is passed on to the public. Also, the meaning of NAEP scores remains a problem. What is the meaning of a 1- to 3-point change, and how should a 1- to 3-point change be interpreted relative to a 5- to 8-point change? Unless ways can be found to interpret the scaled scores and scaled score differences, it may be safer not to report them. Benchmarking scaled scores could be helpful. For example, if the differences between boys and girls in a subject area is 5 points, then this becomes meaningful for judging the relative size of the differences among ethnic groups or changes in NAEP scores over time. Correctly interpreting the meaning of NAEP scaled scores remains a challenge to NCES and NAGB as we move into the 21st century with a new NAEP design and the need for policymakers, educators, and the public to understand the reports they are given about student achievement.

References

Barron, S. and Koretz, D. (1998). *Interpretation and Use of The NAEP TSA Results* (Final Report). Washington, DC: National Center for Education Statistics.

Hambleton, R.K., Brennan, R.L., Brown, W., Dodd, B., Forsyth, R.A., Mehrens, W.A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, W.J., and Zwick, R. (2000). A response to "Setting Reasonable and Useful Performance Standards" in the National Academy of Sciences' *Grading the Nation's Report Card. Educational Measurement: Issues and Practice, 19* (2), 5–13.

Hambleton, R.K. and Slater, S.C. (1994). *Are NAEP Executive Summary Reports Understandable to Policymakers and Educators?* (Final Report). Washington, DC: National Center for Education Statistics.

Hambleton, R.K. and Smith, T. (2000). *An evaluation of the general/public 1996 NAEP Science Reports* (Laboratory of Psychometric and Evaluative Research Report No. 361). Amherst, MA: University of Massachusetts, School of Education.

Jaeger, R.M. (1998). *Reporting the Results of the National Assessment of Educational Progress* (NAEP Validity Studies). Washington, DC: American Institutes for Research.

Kirsch, I.S., Jungeblut, A., Jenkins, L., and Kolstad, A. (1993). *Adult Literacy in America* (Final Report). Washington, DC: National Center for Education Statistics.

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

Messick, S., Beaton, A., and Lord, F.M. (1983). *A New Design for a New Era* (NAEP Report 83–1). Princeton, NJ: Educational Testing Service.

Pellegrino, J.W., Jones, L.R., and Mitchell, K.J. (1998). *Grading the Nation's Report Card*. Washington, DC: National Academy Press.

Wainer, H., Hambleton, R.K., and Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36* (4), 301–335.

Appendix A

Media Study Coding Sheet

_____

Year: __ 1990 __ 1992 __ 1994 __ 1996 __1998
Subject: __ Math __ Reading __ Geography __ History __ Science __ Writing __ Civics
Title of Article: _____

Newspaper: _____ **Date of Article:** ____/____/_____

Paper size: ___ Large ___ Medium ___ Small ___ Other/Not Sure _____

Length of article (words): _____ **Graphics?** _____

1. How are newspapers reporting student achievement?

Report Achievement in terms of . . .		No Mention	Mention	Main Focus	Comments
1.	Standards (B/P/A)	_____	_____	_____	_____
2.	Scaled Scores	_____	_____	_____	_____
3.	National Results	_____	_____	_____	_____
4.	State-by-State	_____	_____	_____	_____
5.	State vs. National	_____	_____	_____	_____
6.	Change over time	_____	_____	_____	_____
7.	Sex	_____	_____	_____	_____
8.	Race	_____	_____	_____	_____
9.	SES	_____	_____	_____	_____
10.	Parents' Education	_____	_____	_____	_____
11.	Interaction (S x R)	_____	_____	_____	_____
12.	Curriculum	_____	_____	_____	_____
13.	NAEP to other tests	_____	_____	_____	_____
14.	Limitations of NAEP	_____	_____	_____	_____
15.	Multiple Subjects	_____	_____	_____	_____
16.	Good Quote	_____	_____	_____	_____

Did the author define achievement levels? ___ Yes ___ No _____

Did the author discuss the achievement levels? ___ Yes ___ No _____

What is the article's main point or focus? _____

_____

2. Identify Misinterpretations: Did the author <u>incorrectly</u> report . . .

<u>Numbers/figures</u>		<u>Comments</u>
% of students in groups	___Yes ___ No	_____
Other numbers	___Yes ___ No	_____

<u>Interpretations</u>

Info about achievement levels ___ Yes ___ No _____

Relationship between variables and achievement ___ Yes ___ No _____

Overall accuracy rating of article? ___ (atrocious) ___ (poor) ___ (good) ___ (excellent)

Example of error: _____

_____

Comments: _____

_____

_____

_____

3. Additional Comments: Note shifts in reporting and how size of the newspaper affects the:

(1) material, (2) length, and (3) the emphasis. Did paper use NAEP conclusions or own?

_____

_____

_____

_____

_____

_____

SECTION 7

Looking at Achievement Levels

W. James Popham University of California at Los Angeles

November 2000

Looking at Achievement Levels[1]

W. James Popham

Subsequent to its establishment by the U.S. Congress in 1988, the National Assessment Governing Board (NAGB) has carried out a variety of congressionally dictated policymaking responsibilities related to the National Assessment of Educational Progress (NAEP). One of these responsibilities (as set forth in the authorizing statute, P.L. 100–297) is "identifying appropriate achievement goals for each age and grade in each subject area to be tested under the National Assessment."

Commencing in 1990, NAGB began establishing "appropriate achievement goals" in the form of achievement levels for student performance in each content area tested. Initially, three levels of quality were used, namely, Basic, Proficient, and Advanced. Then, in 1993 the National Center for Educational Statistics (NCES), the governmental agency responsible for the conduct of NAEP, added a Below Basic category so that NAEP scores would sum up to 100 percent. Thus, the four categories now used in reporting NAEP results are, from high to low, Advanced, Proficient, Basic, and Below Basic.

In the numerous NAEP assessments carried out during the 1990s, with few exceptions, test results reveal an exceedingly small proportion of scores at the Advanced level, a small portion of scores at the Proficient level, and most of the scores in the Basic and Below Basic categories. Generally speaking, about 6% are classified as Advanced; about 19% are classified as Proficient; about 35% are classified as Basic; and about 40% are classified as Below Basic (except in writing, where about 15% of performances are judged to be Below Basic).

A Secretary-Spawned Analysis

During a regularly scheduled meeting of NAGB on November 19, 1999, in Washington, D.C., Secretary of Education Richard Riley made a brief visit to the Board. Having been a founding member of NAGB himself and having served on the Board for 4 years before becoming U.S. Secretary of Education, Riley was well acquainted with the origins and applications of the achievement levels used for NAEP reporting.

During the November 1999 meeting, Secretary Riley asked the Board "to take a look at the current four achievement levels, Below Basic, Basic, Proficient, and Advanced." His reason for this suggestion was his belief that " . . . they're not as useful as I would hope they could be in terms of a person making public policy, whether it's a Governor, or a secretary, or superintendent, whatever. They could help, I think a lot more, convey where improvement is taking place or not taking place and where movement is happening."

[1] An analysis carried out for the National Assessment Governing Board. I am indebted to Mary Lyn Bourque of the NAGB staff who supplied a hoard of pivotal documents and who remained serene despite my seemingly endless questions to her for information about the archived activities of NAEP and NAGB.

Riley continued by noting that, in the last NAEP report about writing, "a great many young people were at the Basic level, a very broad band. And the NAEP report was not able to convey the fact that a lot of these young people were almost very close to Proficient and were significantly higher in that broad band than people at the bottom of Basic."

Riley encouraged Board members "to think about creating perhaps a new category, because the Basic band is so broad. The Basic category and the Below Basic category are very board and then the advanced category is so very narrow that it's hardly useful." The Secretary contended that "the effect, really now, is that we seem to have three categories instead of four and I really think it would be helpful if we had five categories, more categories, using the same numbers, of course."

Riley concluded his remarks regarding NAEP achievement levels by asserting that "the more we can convey to the American people that, yes, we have high standards and none of us wants to toy with that. Challenging, testing high standards is really an overall purpose of NAGB. But, yes, also we're measuring improvement or the lack thereof in a useful way."[2]

One response to Secretary Riley's suggestion was the commissioning, by NAGB's Executive Director, Roy Truby, of an external analysis that led to the current report. The report's author was to provide an "outsider's perspective" regarding the achievement level issue raised by the Secretary.[3]

---

[2] Remarks by U.S. Secretary of Education, Richard Riley, during the November 19, 1999 meeting of the National Assessment Governing Board, Washington, DC.

[3] I wish to affirm that with respect to the endeavors of NAEP and NAGB, I am a *bona fide* outsider. Through the years, dating back to Ralph Tyler's mid-1960s role in devising NAEP, I have regarded NAEP's enterprise in the same way I regard hockey, with thinly veiled disinterest. Now I know that a National Hockey League exists and I realize that many folks seem to live and die on the basis of their favorite team's hockey scores. But I just can't get very excited about hockey. It's a nice game, and I'm glad some people relish it, but I am not among their numbers.

I have always felt similarly about NAEP. It seems like a laudable assessment enterprise, and I'm glad that many of my friends and colleagues are often caught up in National Assessment's intricacies. I'm also sure that NAEP results can be used by policymakers. But my personal interest in assessment, throughout my career, has been in the *instructional* dividends of testing. And, to be frank, I haven't seen NAEP as a formidable contributor to instructional improvements in our Nation's schools. Nor, for that matter, have I sensed that many NAGB members were all that concerned about the instructional yield of NAEP. To them, or so it seemed to me, NAEP was intended to inform educational policymakers, not help teachers make decisions about tomorrow's lessons.

Accordingly, like hockey, I've largely let NAEP go its way while I've gone mine. So, when I accepted Roy Truby's invitation to undertake this assignment, I found that I had to do a substantial amount of background reading. Even before undertaking that reading assignment, without prompting, I could accurately spell NAEP, NAGB, and NCES. But I didn't really understand who was to do what and how it was to be done. So, in preparing to tackle this current assignment, I've learned a good deal about NAEP and how it works. Yet, I am far from being truly knowledgeable about NAEP's nuances or NAGB's preferences. I am, still, a rank outsider with respect to the issues being considered in this analysis. I discovered, however, that the background reading about National Assessment was quite informative. I find myself wondering what would happen if, for some unforeseen reason, I had to read that much about hockey. Perhaps I would find out whether a hockey rink's "blue lines" are really there for any purpose other than aesthetic!

A Preview of This Report's Components

When I was growing up, a "preview" was a short filmed sequence we saw at the motion picture theater, the purpose of which was to advertise a forthcoming film. Such previews are now typically called "trailers," although I have never figured out why something that *trails*, that is, follows, can properly fulfill an *in-advance* advertising role for a forthcoming film. Nonetheless, when sitting in a movie theater, or when reading a report such as this, it seems that most people find it useful to "know what's coming." Accordingly, I'll now provide a brief trailer/preview (choose one) of the chief sections to be found in the remainder of this report.

First, I will attempt to describe how it was that the four current NAEP achievement descriptors came into existence. Having done so, I'll try to discern what might have been motivating those who carved out those descriptors. Next, I will describe what seem to be the motivations for maintaining or abandoning the *status quo* with respect to NAGB's achievement levels. (This analysis will definitely *not* deal with the virtues of the technical procedures currently used to establish achievement levels for the NAEP tests, that question already having been vigorously probed by others on more than one occasion.) Next, I will isolate and discuss the merits/demerits of several likely action-options regarding the achievement levels issue. Finally, I will offer an outsider's recommended course of action for NAGB regarding this very perplexing achievement level issue.

Retracing the Tracks Leading to NAGB's Current Achievement Levels

Let's look, briefly, at the major *documented* events that have led to NAEP's current achievement levels. It is quite likely, of course, that a host of undocumented deliberations also contributed to the evolution of the NAEP achievement levels, but a focus on the chief written documents underlying the process should provide a reasonable picture of how the Board moved from a legislative mandate to its current achievement level policy.

An Early Staff Paper

Soon after its establishment, NAGB initiated its deliberations about how to carry out the Board's legislative mandate to identify "appropriate achievement goals" for NAEP. The first written document devoted to that issue appears to be a December 8, 1989, *Staff Paper on Setting Goals for the National Assessment* (NAGB, 1989). This 20-page, double-spaced analysis dealt with such topics as the legal basis for NAGB's goal-setting efforts, the "case" for standards, and the relationship of NAGB's projected goals to national as well as international goals. The December 1989 staff paper also explored a number of procedural questions such as where the Board should begin, what the goals should look like, and the nature of the goal-setting process itself.

The staff paper also contained a 6-page technical appendix dealing with standard-setting procedures in which it was concluded that "the Angoff methodology is clearly the methodology of choice" (NAGB, 1989; p. 15). Of most relevance to the current analysis, however, was the NAGB staff's conclusion that the Board should opt for a *single* grade-level standard (for all students in a subject area) when establishing appropriate achievement goals. The following paragraph from that staff paper (NAGB, 1989; p. 9) conveys the staff's thinking on this issue.

> The staff struggled with the number of standards that should be set for each grade level. The case for a single standard for each grade is based on the conviction that there is a core of learning in each field that every student ought to master. The case for two levels accepts the assertion that there ought to be a common core of learning, but says that superior performance also ought to be recognized. However, in the final analysis, this paper is premised on a single "universal" grade-level goal for all students in each subject area.

A Public Forum

Although NAGB had, in December 1989, adopted a resolution approving "in concept" the general plan described in its staff's paper, no Board decision was made at that time about the number of achievement levels to be set. In that December resolution, however, the Board did reveal a likely preference about the nature of the achievement goals that it hoped to establish, namely, "grade-level goals that represent solid academic performance, not minimum skills, and which reasonably represent the levels of achievement which all students ought to attain" (NAGB, 1990a).

Next, on January 25, 1990, a full-day forum was staged by NAGB to allow public comment regarding the Board's plan to set NAEP grade-level goals. In addition to a "vigorous" 7-hour debate regarding the plan, more than a dozen individuals and organizations submitted written statements (NAGB, 1990a; p. 1). Focusing here only on illustrative comments relevant to the achievement level issue, Michael Cohen of the National Governors' Association urged that targets be set for three groups of students, that is, the "top, middle, and lowest achieving students." Albert Shanker, President of the American Federation of Teachers, expressed concerns about the establishment of only one standard of satisfactory grade-level performance (as had been suggested in the 1989 NAGB staff paper) because he feared that educators might concentrate their instructional attention on students near that performance level and "not bother" with students well above or well below that single level. The Council of Chief State Schools Officers submitted a written statement, read at the forum, urging NAGB to establish not one but multiple levels using as descriptors "such terms as basic, adept, and advanced." Those submitting their views during the January 25 forum, in person or in written form, offered comments about various aspects of the Board's "approved in concept" plan to establish achievement goals. Yet, in a February 1990 NAGB Bulletin (NAGB, 1990a; pp. 1–6), the aforementioned views were the only documented comments bearing directly on the number of achievement levels that NAGB should adopt.

A Joint Committee Meeting

Almost 1 month after NAGB's public forum, a February 19, 1990, joint meeting of the Board's Technical Methodology Committee and its Analysis, Reporting, and Dissemination Committee (that is, the TM/ARD Committees) took place. Considerable attention was given to the goal-setting process and to the appropriate number of levels suitable for the description of achievement goals for each age, grade, and subject. (NAGB, 1990b).

An initial, 1-page summary of the TM/ARD Committees' conclusions indicated they agreed that the Board should "establish two performance levels at each grade. Possible terms were suggested: 'essential' for the lower level and 'proficient' for the upper one. The levels must be carefully described in terms of test content" (NAGB, 1990; p. i).

Although the summary of the joint committee meeting calls for *two* levels, in the body of the report itself there appears to be an anomalous endorsement of *three* levels, namely, "basic, adept, and advanced" (NAGB, 1990; p. 11). This apparent contradiction may have arisen from the nature of the joint TM/ARD two-committee deliberations or, perhaps, because the 1-page summary of the 13-page report failed to accurately reflect the committee members' preferences regarding an appropriate number of achievement levels.

Establishment of Three Achievement Levels

During a May 11, 1990, meeting in Washington, D.C., NAGB's members unanimously approved the establishment of three achievement levels as well as the technical procedures for identifying those levels at each grade level tested (NAGB, 1990c). Because of the centrality of this important 1990 policy-setting Board action to the achievement level issue being considered here, that report's description of the three achievement levels to be used for reporting NAEP results is presented in full below (NAGB, 1990c, p. 5).

> The central level will be called Proficient. It will represent solid academic performance for each grade tested—4, 8, and 12—and reflect a consensus that students reaching such a level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. At grade 12 the Proficient level will encompass a body of subject-matter knowledge and analytical skills, of cultural literacy and insight, that all high school graduates should have for democratic citizenship, responsible adulthood, and productive work.
>
> There will be one higher level, called Advanced, signifying superior performance beyond proficient grade-level mastery at grades 4, 8, and 12. For 12th grade the Advanced level will show readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement. As data become available, it may be based in part on international comparisons of academic achievement and may also be related to Advanced Placement and other college placement exams.
>
> There will be one level below proficient, called Basic, denoting partial mastery of the knowledge and skills that are fundamental for proficient work at each grade—4, 8, and 12. For 12th grade this will be higher than minimum competency skills (which normally are taught in elementary and junior high schools) and will cover significant elements of standard high school-level work.

The Board promised that the content of each subject area tested by NAEP would support these three achievement levels, and that there would be clear distinctions among the three levels. As the May 1990 report amplified its view of the three levels, it is possible to arrive at a clearer

understanding about how the Board's members were conceiving of the three-level classification scheme (NAGB, 1990c; p. 6).

> These benchmarks will permit States and the Nation to see what proportion of students have reached very high levels of achievement on NAEP exams; strong, acceptable levels; and levels of partial mastery. Thus, it will provide a measure and incentive to improve the learning of all segments of the distribution—bottom, middle, and top.
>
> The framework of three achievement levels at each grade is not a warrant for tracking. Indeed, the NAEP tests and the achievement levels based on them will help to ensure that all students attain competency in challenging subject matter.
>
> The proposed achievement levels will define levels of learning tied to a common core of knowledge and skills that ought to be available to all students, regardless of family income, ethnic background, region, or type of community. The achievement goals on the National Assessment will serve to underscore the point that American schools ought not to water down what they teach the poor and beef up what they offer the more affluent.

The principles of the Board's anticipated technical procedures for establishing the three achievement levels were also spelled out, in considerable detail, in the May 1990 report. Although it appears that a number of the technical procedures recommended in 1990 were, quite understandably, modified when they were subsequently implemented, the technical template for determination of three NAEP achievement levels was clearly set forth in NAGB's May 1990 report.

There is another theme often stressed in the 1990 report, namely, that the establishment of achievement level goals for NAEP did not constitute the creation of a federally mandated curriculum. Board members clearly wished to make the National Assessment more useful to "parents and policymakers as a measure of performance of American education and perhaps as an inducement to higher achievement." Yet, writers of the report quickly add that "the achievement levels will be benchmarks, points for judgment and encouragement, not edicts or commands" (NAGB, 1990c; p. 13).

The Birth of "Below Basic"

Before the introduction of NAGB's three 1990-approved achievement levels, NAEP results had been reported using briefly described numerical levels of proficiency along with the percentage of students who had scored "at or above" each of these levels at each grade tested. Thus, for example, national mathematics results might be reported at proficiencies of 200, 250, 300, and 350, each of which was accompanied by a terse description of the skills a student at that numerical level could perform. To illustrate, a 200 level for mathematics in 1990 was described as "Simple Additive Reasoning and Problem Solving with Whole Numbers." Then, the percentages of students who were "at or above" this level were reported for the grades tested. For example, in mathematics, at "average proficiency" level 250 (Simple Multiplicative Reasoning and Two-Step Problem Solving), the 1990 percentages of students "at or above" this level were, respectively,

11% for grade 4, 67% for grade 8, and 91% for grade 12. This same "at or above" approach to score reporting was initially carried over to the three NAGB-defined achievement levels.

When NCES began in 1993 to publish reports of the proportions of children scoring at the three NAGB-approved achievement levels for the 1992 mathematics test, the reporting scheme relied on "at or above" labels. For instance, at grade 4 in the 1992 NAEP reading test, 27% of the Nation's test takers scored "at or above Proficient" (thus embracing both those who were classified as Advanced and those who were classified as Basic). Yet, when NCES staff released percentages "at or above Basic" in grade 4 reading, only 60% of the scores were actually reported. To account for the missing 40%, NCES created a Below Basic column, thereby allowing NAEP reporting categories to sum up to a universally cherished 100%.

NAGB may not have been formally consulted regarding this decision, which was understandable in view of NCES' exclusive reporting responsibilities for NAEP. Nor did the Board thereupon endorse Below Basic as an achievement level. This appears to stem from some Board members' views that the NAGB-sanctioned achievement levels should not represent mere reporting categories but, rather, should be considered achievement *goals* as set forth in the authorizing legislation creating NAGB (that is, when it was indicated, by law, that the Board should identify "appropriate achievement goals").

Thus, since the early 1990s, there has been either confusion, disagreement or, more often, both confusion and disagreement about the Below Basic category. Is it even appropriate to conceptualize a Below Basic performance category in the same manner that one thinks about a Proficient performance category? NAGB, in a November 20, 1993, policy statement (NAGB, 1993), still refers to three achievement levels, but does not rule out the possible interpretation of these achievement levels as achievement goals."

The 1994 Reauthorization

The *Improving America's Schools Act of 1994* (P.L. 103–382) reauthorized NAGB's overall policy role with respect to NAEP, but stated in Section (e)(1) that the Board "shall develop appropriate performance levels for each age and grade in each subject area to be tested under the National Assessment." The reauthorization also indicated that these performance levels should be "reasonable, valid, and informative to the public." Moreover, the performance levels ought to be "updated as appropriate."

The Board, for the sake of continuity, decided that the "performance levels" called for in the 1994 reauthorization would continue to be called "achievement levels." Moreover, in a March 4, 1995, policy statement (NAGB, 1995), NAGB reaffirmed that the achievement levels would be regarded as "expectations which stipulate *what students should know and be able to do* (emphasis in original) at each grade level and in each content area measured by NAEP."

The Board also foresaw that NAEP's achievement levels would be helpful in interpreting the meaning of the National Educational Goals that had been codified in the 1994 National Education Goals 2000 legislation, particularly those goals dealing with students' mastery of academic subject matter.

In the March 1995 policy statement (NAGB, 1995; p. 2), NAGB provided more streamlined descriptions of the three NAEP achievement levels:

- **Proficient:** 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.

- **Basic:** This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.

- **Advanced:** This level signifies superior performance beyond Proficient.

A Call for a National Test

In his February 1997 State of the Union Address before Congress, President Bill Clinton called for the creation of national achievement tests in reading and mathematics. Shortly thereafter, the U.S. Department of Education initiated development of a Voluntary National Test (VNT) focused on grade 4 reading and grade 8 mathematics. In August 1997 the Department of Education entered into contracts with external agencies to initiate developmental work on the VNT. One of those external contractors, the American Institutes for Research (AIR), was to coordinate the efforts of other VNT subcontractors.

NAGB became involved with the VNT only during the fall 1997 appropriation negotiations, which led to the resulting legislation in November calling for NAGB to be given "exclusive authority" over the AIR contract. As AIR and its subcontractors moved forward with the development of the VNT, one of its early requirements was to locate sufficient numbers of students so that the under-development VNT items could be field tested. The Council of the Great City Schools, a national consortium of 50 of the Nation's largest urban school systems, was approached by VNT developers to secure field-test sites. The response by the Council was supportive of the VNT and its mission, but Council members expressed serious concerns about the adequacy of the three NAGB-approved achievement levels.

In an appearance before NAGB during its March 7–9, 1998, meeting in Arlington, Virginia, Michael Casserly, Executive Director of the Council of the Great City Schools, registered his organization's support for the chief thrust of the VNT. He said that the Council had agreed to have its students take the proposed VNT because "we wanted to make it crystal clear to the Nation that urban school leaders want and expect the highest standards for our students."

Having expressed support for the VNT efforts, however, Casserly identified the Council's concerns about the three achievement levels apt to be used for reporting VNT results:

> . . . as I understand it, the Voluntary National Test will use the same framework in skill levels as NAEP, Basic, Proficient, and Advanced. The Council generally agrees with this, but remains worried about the lack of information provided to

> students and parents and teachers for individuals who score below the basic level. Recent NAEP results indicate that around 35 percent of students nationally score below the Basic level.
>
> And the recent *Education Week* report, Quality Counts, shows that an average 50 percent, some say between 30 and 90 percent of urban public school students nationally, may not reach that level. The result could easily be that a substantial number of individual students taking the test in urban areas would have no data on their status, defeating the purpose of the student-by-student reporting, and making it more difficult to spur any meaningful community concern . . . .
>
> [This] could mean that we have volunteered for an exam on which we could get almost no results. Let's take an extreme example. If 60 percent of the 4th graders in a particular city are English language learners and cannot respond substantively to the reading test, and some 50 percent of the remaining students do not hit the Basic level, then we could be left where only 20 percent of the students are provided meaningful reports on their individual status on the standards. I don't think this is what all of us had in mind.
>
> Unfortunately, the Council does not yet have a solution to this problem, but we urge the Board to work with us and others to resolve it. We need to provide students Below Basic with some measure of what they can do and what they cannot do without lowering the standard itself.

Even though the concerns about achievement levels from the Council of the Great City Schools arose from apprehension regarding the VNT, a test whose future is currently uncertain, those same concerns have begun to enter the dialog about the three NAEP levels themselves. AIR, still authorized to develop VNT items, has urged NAGB to create more descriptive information and exemplar responses for the Below Basic nonlevel. As current NAGB policy stands, however, "Below Basic" has not been formally designated as an endorsed achievement level for reporting NAEP results.

Two Rounds of NAGB Public Meetings

During March 29–April 12, 1999, NAGB sponsored four 1-day public hearings throughout the Nation on the VNT's purpose, intended use, definition of the term "voluntary," and reporting. Because the NAEP reporting levels were to be used if the VNT became operational, comments during the four 1-day meetings about reporting have some relevance to the issue under consideration here.

The public was specifically invited to comment on, among other issues, the reporting of results for students whose performance is Below Basic. Several witnesses expressed the view that more useful information should be provided regarding such students. At least one witness recommended the creation of a system that would promote better instructional practices for use with Below Basic students (NAGB, 1999).

A somewhat similar series of four 2-hour discussion groups was sponsored by NAGB between September 9 and December 9, 1999, to explore public perceptions of the NAEP achievement levels. The report of those meetings, a good deal of which bears on the focus of this achievement levels analysis, is available in a separate report (NAGB, 2000).

It is clear that NAGB recognizes the significance of concerns regarding the NAEP achievement levels and, as a consequence, has initiated a number of activities to provide the Board's members with information needed to retain or modify NAGB's current achievement levels policy. Indeed, a meeting of the Board's Achievement Levels Committee has been scheduled for June 23–24 in Utah to consider NAGB policy in light of a number of reports commissioned by the Board to study various aspects of the achievement levels issue.

When Goals Become Levels

Looking back over the more than 10-year NAGB struggle with achievement levels, one is struck by the potential for conceptual confusion stemming from the sometimes interchangeable use of the terms "goals" and "levels." When NAGB was born, its authorizing legislation called for the Board to identify "appropriate achievement goals for each age and grade in each subject area to be tested under the National Assessment." Now a *goal*, as most people understand the term, is precisely what the dictionary says a goal is, namely, "the result or achievement to which effort is directed; aim; end." So, if NAGB was supposed to establish achievement goals for NAEP, it ought to be in the business of setting forth *aims* for NAEP subjects and grade levels. Perhaps this is why the original December 1989 NAGB staff paper recommended a single achievement level because a solitary level of student performance would, indeed, represent a goal.

But when NAGB's May 1990 policy established *levels* to describe NAEP achievement, it really departed from a strict conception of "goals as aims." I am not disputing the wisdom of the Board's May 1990 policy. Indeed, it may have been precisely the kind of clarified reporting system that NAEP needed. But descriptive levels, if we wish to use language precisely, are not equivalent to goals.

Even in that important 1990 policy document, an ambivalence existed about whether NAGB had simply set up a suitable framework for describing students' NAEP performances or, instead, had satisfied the "letter" of the authorizing law by carving out achievement *goals.* In a few sentences appearing just before the report set forth the three Board-approved achievement levels, the report's authors engaged in "goal-talk" even as they refer to a *measurement* mission in the form of NAEP's "yardstick" role.

> Defining what performances ought to be—and providing strong justification for the judgment used in making these definitions will greatly enhance NAEP's central function as a yardstick of educational achievement (NAGB, 1990c; p. 5).

In retracing the events and policies leading to the achievement level questions currently facing NAGB members, it is possible to detect foreshadows in the influential report on *The Nation's Report Card*, carried out under the leadership of Lamar Alexander and H. Thomas James (Alexander and James, 1987). In that report the establishment of a NAGB-like group (an

"Educational Assessment Council") was recommended. One of the responsibilities of that assessment policymaking group, according to the report, would be "identifying feasible achievement goals for each of the age and grade levels to be tested" (Alexander and James, 1987; p. 32). (The language of the Alexander-James Study Group's report, it will be noted, is essentially identical to that employed in the legislation authorizing NAGB.) So, in the study group's report, there is a recommendation that a NAGB-like council identify achievement *goals* for the National Assessment.

Yet, in a companion review of the Alexander-James Study Group's report, a committee of the National Academy of Education offered a recommendation that "NAEP use descriptive classifications as its principal reporting scheme in future assessments" (National Academy of Education, 1988). The review committee, chaired by Robert Glaser, continued that:

> For each content area NAEP should articulate clear descriptions of performance levels, descriptions that might be analogous to such craft rankings as novice, journeyman, highly competent, and expert. Descriptions of this kind would be extremely useful to educators, legislators, and an informed public (National Academy of Education, 1988; p. 38).

Thus, even in the chief documents that helped spawn NAGB, we encounter contrasting preferences. One set of players wants NAEP to be accompanied by achievement goals. One set of players seeks more meaningful descriptors of students' NAEP performances.

The 1990 NAGB report's "two-for-one functions" position seems to presage the following decade's continuing concerns about NAEP's achievement levels. If those levels were being explicitly used as surrogate goal statements, then it would seem there should be strong NAGB endorsement of *one* of the levels as the true *target*. And this seems to have been done in that the Board has periodically contended the *Proficient* level (and, of course, as many *Advanced* performances as can be had) is where the Board's aspirations really lie.[4]

But many individuals, for example, members of the Council of the Great City Schools, seem to regard the achievement levels less as goals than as descriptive levels—levels they think will not be useful in describing the performances of many students in urban school settings (Casserly, 1998).

And even when the reauthorization of NAEP occurred in 1994, the legislation's switch from "goals" to "performance levels" did not erase this definitional confusion. The legislation directed the Board to "develop appropriate student performance levels for each age and grade in each subject area to be tested under the National Assessment" (NAGB, 1995). For some individuals, performance levels are only descriptive categories. Yet, because the reauthorization refers to "*appropriate* performance levels," it is also reasonable to regard such levels as goals. Again, the goals versus descriptors confusion lingers.

[4] See, for example, the National Assessment Governing Board (April 27, 1995) news release by William T. Randall, NAGB Chairman, describing the 1994 NAEP reading results. In that release, Randall contended NAGB regards the Proficient level as "one we believe all American students should reach."

What I am suggesting is that at least a segment of any current disagreements about NAEP achievement levels arises because some people consider such levels to be goals, some people consider them to be descriptive labels, and some schizoid-free folks blithely consider them to be both.

Recognition of this definitional difficulty does not resolve it. Yet, as NAGB members wrestle with their choices about how best to describe students' NAEP performances, they will find it useful to distinguish, at least in their own minds, whether the *goals* or the *descriptors* function of the levels is under consideration at that moment.

If NAGB moves *fully* toward a "levels-as-descriptors" versus "levels-as-goals" position, then a number of policy possibilities arise. For instance, if the Proficient level is, in reality, no longer a goal—only a descriptive category—then less stringent performance expectations might be established for that level because it is, indeed, only a label and not a target. If, on the other hand, NAGB *completely* embraces the "levels-as-goals" position, then it might be possible to maintain the Proficient level as a target, but create a somewhat parallel but not coterminous set of purely descriptive labels.

The current hybridized use of NAEP's achievement levels, even if better understood in historical context, appears to be fostering not only confusion but, on the part of many, genuine distress. It may well be time for NAGB to take an unequivocal definitional position on this issue, then pursue that position's implications either via vigorous clarification or, perhaps, policy alterations.

Motives for Modifying/Retaining NAGB's Achievement Levels Policy

I turn, now, to the reasons that seem to underlie a fairly widespread concern about the virtues of NAGB's current achievement levels policy. I'll first consider the reasons that some people want to alter the current achievement levels policy, then deal with the reasons that some would like to see the policy stay as is.[5]

Motives for Modification

A variety of reasons have been forwarded by those who wish to see the Board's current achievement levels policy altered. Several of these, of course, are embodied in the comments of

[5] I cannot resist the recounting of a somewhat eerie coincidence regarding this report. Actually, I had planned to author the report in two separate sections, the first part focusing on the content treated up to this point, and the second part dealing with the content to be considered hereafter. This allowed NAGB members to review a draft version of the report's first section during their March 3–4, 2000, meeting in Honolulu and to offer suggestions regarding how the achievement-levels issues might be addressed in the report's final pages. I appreciate the useful suggestions of NAGB members during that March meeting. Thus, when on March 17 I was able to begin writing the report's second section, I was surprised to read an essay in the *Honolulu Advertiser*, Hawaii's leading newspaper, in which the writer, Cliff Slater, deplored the poor quality of the State's schools. One of the arguments he cited to support his argument was a national report about NAEP results asserting that in Hawaii "the proportion of 4th graders reading at the Proficient level [is] now the lowest percentage in the Nation." He then went on to suggest changes in Hawaii's educational system based on this negative evidence. If I ever needed a reminder that people can be influenced by the way that NAEP's results are reported, there it was in my morning paper!

Secretary Riley (see pp. 159–160 of this section) Briefly, then, five of the chief dissatisfactions with the three-tiered achievement level structure will now be identified and discussed.

A Failure To Provide Policymakers With Sufficiently Useful Information. Secretary Riley observed in his November 1999 remarks to the Board that the four achievement levels (including, of course, Below Basic) simply did not supply educational policymakers with sufficiently useful information. In part, this is attributable to the lack of within-level subcategories so that Secretary Riley and others can see, as he said, "where improvement is taking place or not taking place."

Thus, conceiving of the levels as reporting categories rather than goals, some analysts believe there is not sufficient within-level differentiation to allow policymakers to discern whether, over time, improvements in student performances on NAEP are or aren't taking place.

As Edward Haertel pointed out during the NAGB March meeting in Honolulu, the differentiating deficits of the achievement levels becomes especially salient if one applies the current NAEP achievement levels to the description of an individual student's performance (as is proposed for the VNT) rather than to the description of aggregated data (as would be true when policymakers review NAEP performances for a group of students). Yet, as Haertel observed, those same difficulties with undifferentiated within-level results can prove genuinely troublesome even to policymakers who employ aggregated NAEP results.[6]

The Excessive Breadth of the Basic and Below Basic Reporting Categories. A related difficulty that some see in the NAEP reporting categories is that at least two of them are simply too broad. Because, in the main, about three-fourths of students' NAEP performances are classified in the Basic and Below-Basic categories, it can be argued that these two categories are excessively broad. If there were within-level performance subcategories, of course, this "too broad" attack might have less cogency. However, even without arguing the merits of within-level differentiation, a four-category descriptive scheme in which two categories capture 75% of the scores while the other two categories account for only 25% of the scores appears, *a priori*, to be imbalanced at least in part because of certain categories' excessive breadth.

Advanced and Proficient Levels That Are Unrealistically High. Because so few students are classified in the Advanced and Proficient levels, that is, respectively, 6% and 19%, some critics contend that both of these achievement levels were set at an unwarrantedly high level. This criticism, of course, becomes more serious if one conceives of the achievement levels as descriptive categories rather than goals. However, even if one thinks of the Proficient level as the central goal that should be sought, one can still argue that the Proficient and Advanced levels have been set at levels so high that, realistically, too few students can attain those unwarrantedly high levels.

A Goal-Based Reporting Scheme That Makes Educators Appear Ineffectual. Related to the criticism that the Proficient and Advanced levels have been set too high, one resulting perception is that American educators are not sufficiently effective. If so few students achieve Proficient-

[6] Haertel, Edward H., remarks made during the NAGB March 3, 2000 meeting, Honolulu, Hawaii and in a subsequent personal communication with the writer.

level (and above) status, this suggests that the Nation's educators aren't doing a very good instructional job. Negative perceptions of the effectiveness of American educators can lead to a host of educational policies that, in both the short and long term, are apt to have an adverse effect on American schools.

In a similar vein, at the State level it often turns out that students' performances on State-devised accountability tests look much better than do students' performances on NAEP. For the educators in such States, then, the State-devised tests become regarded by the public as "homegrown softies" while the "real" national tests expose a State's educators as ineffectual. Is it any wonder that those who must decide whether a State takes part in NAEP regard such participation warily. [7] Even if a State's educators have done an honest and careful job in setting up their own statewide assessment system, the stringent achievement levels for NAEP set by NAGB will almost certainly disconfirm what otherwise might be seen as a State's educational success story.

A Goal-Level Label That Renders Lower Performances Unacceptable. The labeling of a goal structure or a set of performance levels is a peril-fraught enterprise. It is far easier to miss the mark when one labels than it is to carve out descriptors satisfying all. I empathize, therefore, with the NAGB architects who, more than a decade ago, chose "proficient" as their target goal level. And this goal level, as the Board has often reiterated, was intended to represent a challenging level of student attainment, one that represents "solid academic performance." Yet, in choosing the label for the desired level as "proficient," the Board thereby rendered all lower levels of performance as *not proficient.* And the large proportions of students earning *not proficient* scores on NAEP tests is just what the news media love to report because negative stories surely attract more attention than positive ones. The use of the "proficient" label for an achievement level that only a quarter of test takers will attain dooms three-quarters of test takers to be regarded as seriously wanting.

Consider, in table 1, the four descriptive categories now used by NAGB alongside the four descriptive categories used to illustrate performance level descriptions in the January 1988 National Academy of Education (NAE) response (National Academy of Education, 1988; p. 38) to the Alexander-James report. The bold-faced descriptive categories are thought, by most people, to be clearly unacceptable.

Table 1. Comparison of Two Sets of Descriptive Labels Used by NAGB and NAE (bold regarded by most as unacceptable)

NAGB Levels	Illustrative NAE Levels
Below Basic	**Novice**
Basic	Journeyman
Proficient	Highly Competent
Advanced	Expert

[7] I have personally heard educational policymakers in two different States indicate that if the achievement levels for NAEP are not somehow altered, they will urge that their States withdraw from NAEP. As one of them informed me emphatically, "As NAEP's reports now exist, it can *only* make this State look bad. Our own test results are good, but NAEP makes us look lousy!"

As is seen in table 1, given the actual performance expectations set for the Proficient and Advanced levels set by NAGB, students who fall in the two lowest categories of the NAGB model are seen as *not proficient*, hence inadequate. In the NAE illustrative categories, a "journeyman" usually carries a positive connotation with it, thus only one category of the NAE descriptors is patently unacceptable.

Thus, one reason for dissatisfaction with the NAGB achievement levels is simply that the descriptor used for the *acceptable* level automatically demeans all lower levels of student performance.

Descriptors and Goals. Harkening back to the earlier discussion of the distinction between viewing NAGB achievement levels as goals versus descriptors, it is apparent that such a distinction bears directly on several of the five motives for modification identified here.

Motives for Retention

In contrast to the sometimes more vocal critics of NAGB's achievement levels policy, there are those who are more than a little unwilling to see the current labels or the rationale underlying them altered in any way.

Continuity's Dividends. A good many people who have followed NAEP's activities through the years are loath to undertake any actions that would reduce the comparability of students' performances over time. Born in the 1960s, NAEP is one of the few enduring educational assessment enterprises in America. If it is tinkered with in any meaningful manner, some fear that the over-time interpretability of NAEP will be markedly diminished.

Softened Performance Standards. Others who are familiar with NAEP, and who value its contributions to State and national policymaking, do not wish to tamper with the achievement levels or labels in any manner that might make the public believe a rigorous assessment system has been softened. Because in its present form NAEP seems to show the Nation's schools to be less effective than desired, it may appear to the public that if any changes in the reporting structure are made, then NAEP results are simply being massaged to make educators *appear* more successful than, in reality, they are.

Status Quo Devotees. It must be recognized, of course, that there are a good many individuals who are simply subscribers to the status quo in whatever form it exists. Such individuals are reluctant to change things for just about any reason, short of those that might be life threatening.

Milder forms of such status-quoism can be found in those individuals who are unwilling to tamper with a NAGB-sired reporting model that has now become so widely adopted by the Nation's educators. According to a recent report by Jeff Nellhaus (Nellhaus, 2000), about half of the Nation's 50 States now employ performance standards that are identical with, or similar to, those promulgated by NAGB. It is easy to understand how someone would be loath to modify a model that now appears to be well accepted by so many American educators.

Retention's Rewards. It is sometimes tempting to identify those who wish to make changes as the "improvers," but to regard those who support the status quo as the "improvers' enemies." Given the mild clamor that sometimes seems to be heard from those who wish to see an overhaul of NAGB's achievement levels policy, it would be easy to dismiss the views of those who want no significant changes as some sort of "status quo squatters." Yet, given the impact of NAGB's decade-old decisions about the nature of NAEP's reporting scheme, the views of those who wish few or no changes in the current NAEP reporting model must be taken seriously. Their "stand-pat" stance is far from silly.

Shortcomings Viewed as Serious

It must be recognized, however, that the motivations of those who wish to alter the nature of NAGB's achievement levels policy are based on serious concerns about what they regard as significant shortcomings in an important assessment system. More than a few of the people who urge meaningful changes in the current achievement levels policy believe that if no major modifications are made, the utility of NAEP-produced data will be seriously diminished. Given the nation's enormous investment in NAEP over the years, an investment in both dollars and intellectual capital, the advocates of alteration believe that such changes must be made or NAEP's significance will be greatly reduced.

So, even though it is possible to consider potential modification options, as will be done in this report, such an activity is far from a mere academic exercise. There are NAEP-knowledgeable policymakers out there who, if the achievement levels are not significantly altered, will certainly urge NAEP's demise.

Modification Options

Having considered the chief motives for modifying or retaining NAGB's current achievement levels policy, I will now focus on five modification options that appear to be likely contenders for change. I will attempt to consider both the strengths as well as the shortcomings of the proposed changes. Accordingly, the alert reader should prepare to encounter an "on the other hand" phrase more than a time or two.

Although I briefly described these and other modification options to NAGB members during an early March meeting of the Board in Honolulu, and although several of the options were discussed by members of the Board, the final five modification options I wish to consider in this report are:

1. Add one or more achievement levels.

2. Divide the current levels into distinguishable, within-level reporting categories.

3. Make Below Basic a NAGB-sanctioned reporting category.

4. Relabel the existing achievement levels, especially Proficient.

5. Lower scale-score ranges associated with one or more achievement levels.

Adding Achievement Levels

One possible change would be to add one or more achievement levels to the four current reporting categories. Proponents of this change, perhaps including Secretary Riley (see p. 160 in this section), believe that the introduction of at least one additional level would make it possible to differentiate more accurately among the roughly 75% of students now classified as Basic or Below Basic. A reporting model with one or two more categories, they believe, will yield more useful information to policymakers because the potential for detecting increases as well as decreases will be augmented.

On the other hand, one or two new reporting categories, especially if they are inserted below the Proficient level, will do nothing to erase the continuing perception that hoards of American students are incapable of performing at a Proficient level. Moreover, an increase in the number of NAEP reporting categories would run counter to the now widely used four-category reporting structures that NAGB's model has stimulated in so many parts of the Nation.

Subdividing Achievement Levels

Those who regard the current achievement levels as too broad, particularly the lowest two reporting categories, have suggested that the current levels can be split into distinguishable within-level categories such as "high" and "low" or even "high," "middle," and "low." The advantage of this option is that it would differentiate more accurately students' locations in an otherwise too-broad reporting category, yet maintain the original four-category reporting model.

Opponents of such a change option, on the other hand, argue that at present there are insufficient numbers of items in the NAEP item pool to make possible any meaningful within-level differentiation, particularly at the lowest and highest ends of the performance distributions.

Moreover, such critics believe that within-level differentiation still doesn't address the concerns of those who view with chagrin the large proportions of students who are scoring well below the Proficient level, hence who will still be viewed by the media as not proficient.

Blessing the Below Basic Category

Another option to change the current NAEP reporting model is to have NAGB more fully sanction the existence of the Below Basic reporting category. This could be accomplished by better describing student performance at this level as well as by offering sufficient illustrative student responses to communicate more clearly what this lowest level of student performance truly looks like.

Although this proposed change would help educators in NAEP low-performing schools get a better fix on the nature of their students' performance levels, it would not seem to satisfy the concerns of groups such as the Council of the Great City Schools who fear that the vast majority

of their students will end up as being classified well below NAEP's Basic level. Better understood inadequacy is, unfortunately, still seen as inadequacy.

Relabeling Levels

A fourth option to alter the current achievement levels model deals directly with nomenclature. Proponents of a relabeling approach believe that if different descriptors were used, especially the label for the Proficient category, then a reporting model could be fashioned so that NAGB's designated goal level could be named something that would not lead to all levels below it being characterized so negatively, as is currently the case with all those below Proficient who are labeled, predictably, as "not proficient."

On the other hand, opponents of this relabeling option believe that the Nation's citizens, and especially its educational policymakers, would readily see through such a transparent attempt to dress up NAEP results in more palatable costumes. If it is true that a rose, regardless of its name, smells sweet, then these critics claim that sauerkraut, even if gift packaged, will still smell sour. Besides, such critics contend, mid-course relabeling will be widely regarded as a blatant instance of NAEP standards softening.

Lowering Scale-Score Ranges

A final recommendation for change springs from those who believe that the original expectations set for the Proficient and Advanced levels were simply too high. Some reduction, perhaps not dramatic, would render the NAEP four-level reporting structure more realistic. The setting of performance standards, such proponents argue, is still, at bottom, a judgmental enterprise. Hence, they believe that the originally selected scale-score ranges are obviously too high, so should be lowered for the top two reporting categories.

On the other hand, those who find this option unacceptable contend that such an obvious lowering of performance standards would present a clear admission to the world that the Nation's much-touted pursuit of *demanding* levels of student performance was little more than public-relations rhetoric. Critics of the "lower-the-required-scores" option will claim that such a crude approach to this issue would forever damage NAEP's credibility because it would be seen as little more than a self-serving education profession's adjust-as-needed yardstick.

Looking Over Options

None of the five options proposed here leaps out as a sterling solution-strategy that will instantly ameliorate the ills some people believe now afflict NAGB's achievement levels policy. Each solution, or so it seems to me, has some virtues that are, distressingly, accompanied by one or more vices. None seems, all by itself, to reach out and say, "Choose me; I am the one."

My personal opinion, however, is that *something must be done* to alter an achievement levels policy which, if not modified, may make NAEP an educational anachronism within a decade or two. I believe NAGB's consideration of this issue must result in some sort of meaningful reformulation of the Board's achievement levels policy.

So, with all sorts of uneasiness, I now offer an outsider's recommendation to NAGB's members about how to modify their current achievement levels policy. It is not a perfect solution that will satisfy all. Given NAEP's history and today's educational realities, I do not think there is an in-waiting solution-strategy that will garner universal acclaim. I'm sure my proposal won't either. That being the case, along the way, I'll try to defend my recommendation.

An Outsider's Recommended Solution-Strategy

The rationale for my recommendation hinges on the belief that NAGB's original position to establish demanding *achievement goals* was absolutely correct. But, in the resulting contamination of "levels as goals" with a "levels as descriptors" perception, what was originally a high-aspiration goal has become a reporting-focused cause for what is, at best, serious confusion or, at worst, an insidious means of discrediting America's educational enterprise. So, I want to see the goals/descriptors confusion clarified unequivocally by NAGB while, at the same time, making significant changes in how the two lowest level NAEP reporting categories are used.

Reaffirmed Goals

In essence, I want NAGB to reaffirm, and broadcast such reaffirmation widely, that the *goal* of American education should be to get increasing numbers of the Nation's children to attain *at least* the "solid academic performance" represented by the Proficient level. Proficient-level mastery (and, of course, Advanced-level mastery if possible) must be the Nation's unmistakable goal, a goal so worthwhile and challenging that citizens should recognize not all students *at this time* have reached that goal *or will rapidly do so.* Nevertheless, in this new century, the Nation's children cannot be satisfied if they achieve less than such a high level of performance, nor can the educators who are charged with getting our students to that challenging level of mastery.

Subcategories Within Two Reporting Categories

To help more and more students achieve proficiency-level status, the two lower levels reporting categories should be divided into distinguishable levels of student performance, ideally, "high," "middle," and "low."[8] NAGB should make clear to all that the differentiation of its two *below-goal* reporting categories represents an unabashed *improvement* tactic, that is, an effort to help more students achieve the proficiency-level *goal.* In other words, I recommend within-level differentiation only in the Basic and Below Basic categories, and only because of a well-publicized commitment by NAGB to help America's students better master the important knowledge and skills measured by NAEP. Students who have not yet reached the Proficient level goal *need improvement.* A within-level set of improvement-facilitative subcategories can help the Nation's educators track such improvements.

[8] I recognize that an insufficiency of NAEP items may currently make a three-split within-level distinction impossible, but I would insist that NAEP's item developers immediately get busy in producing sufficient items so that a three-split, within-level reporting scheme will be possible. A two-split, within-level approach (e.g., "high" and "low) would be better than nothing, but in my view a three-split approach would be immeasurably preferable.

Moreover, the creation of six well-defined and exemplar-illustrated within-level subcategories ("high," "middle," and "low") for each of the Basic and Below Basic categories should be carried out in the context of an *instructional* orientation that provides as much guidance as possible to the educators who are striving to help students reach higher levels of NAEP-assessed achievement.[9]

If the six within-level subcategories of the two lowest NAEP reporting categories can be regarded as steps in a "needs improvement" stairway culminating in the student's attainment of proficiency-level performance, or higher, then as much instructionally facilitative attention as possible can be lavished on the delineation of those subcategories and how they might be addressed instructionally.

A Graphic Representation

This recommendation is represented graphically in figure 1, where it can be seen that the six lowest subcategories (constituting the Basic and Below Basic reporting categories) are all indicative of student performance that needs to be improved. The two highest NAEP reporting categories represent, of course, NAGB-established achievement goals for the Nation.

As I see it, the chief factors of the solution-strategy recommended here are these: A within-level subdivision of the Below Basic and Basic reporting categories. A reaffirmation of the Proficient level as the quality of performance sought from all students. A clarification of all six "needs-improvement" subcategories, complete with descriptive language and exemplars, ideally fashioned with instructional decisionmaking in mind. A major public dissemination effort to familiarize all relevant constituencies with the chief features, and the reasons for those features, of the revised NAEP reporting system.

As is apparent, if the proposed recommendation were to be adopted by NAGB, a substantial outreach educative effort would most certainly need to be carried out so those concerned would see that NAGB's educational aspirations remain constant, but also recognize that a deliberate effort has been made to help those students who need improvement. The heart of that effort will be NAGB's creation of a more fine-grained gradation of student capability within the two below-goal NAEP reporting categories.

In the midst of such an educative effort, of course, great care should be taken to clarify the "levels-as-goals" versus "levels-as-descriptors" distinction previously discussed. In that connection, the Below-Basic level should be clearly sanctioned as a NAGB-endorsed *descriptive* category.

[9] Although I recognize there are some potential perils associated with the perception that NAGB is somehow subtly pushing a national curriculum, I regard the NAEP curricular frameworks as first-rate documents helpful instructionally. It is a shame they are not more widely used in the nation's schools. Perhaps those curricular frameworks could be somehow linked to the six below-goal subcategories I am suggesting.

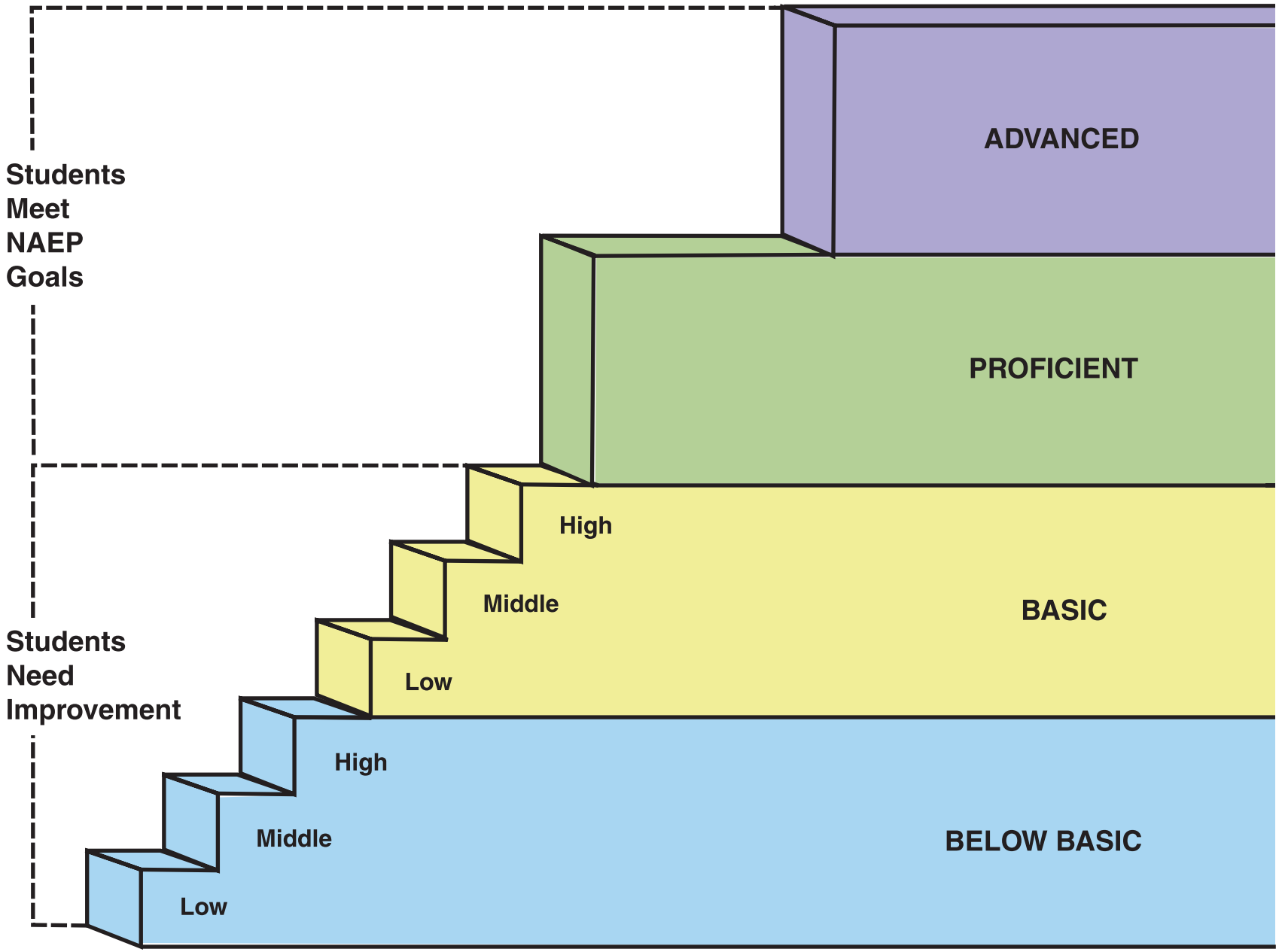


Figure 1. A recommended modification in NAEP reporting categories.

Advantages of the Recommended Modification

The recommended modification would, in my view, accomplish the following:

- It would not alter either the level of NAGB's expectations for American students, nor would it change the labels associated with NAEP's reporting structure. Thus, there would clearly be assessment continuity and no lowering of standards.

- It would make possible what seemed to be the central focus of Secretary Riley's concern, namely, the identification for policymakers of situations where improvement was or wasn't taking place. The newly created six subcategories within Basic and Below Basic would lend themselves ideally to this kind of use by policymakers.

- It would make possible, because of the heightened focus on the six "needs improvement" steps toward goal mastery, for NAGB to provide more instructionally catalytic support for those educators who wish to employ NAEP's curricular frameworks and achievement subcategories in educative approaches that would benefit students.

A Two-Tactic Strategy Required

For this solution-strategy to have any chance of working, two communication tactics must be immediately installed. Both of these depend on NAGB's making a clear, powerful distinction between levels as descriptors and levels as goals. NAGB must strongly and frequently communicate the message that the purpose of the current four reporting categories is definitely to *describe* NAEP performances, but that the Proficient achievement level unequivocally constitutes the high-level *goal* sought for American students.

The first communication tactic, then, is to make clear to all that the *demanding* nature of the Proficient level goal makes it likely, at present or in the immediate future, only a relatively small proportion of U.S. students will attain such a *demanding* level of mastery. NAGB needs to persuade the public that, in the near term, it will be a formidable educational task to get even half of the Nation's students up to that challenging level of performance. The challenging cognitive demands embodied in NAEP's goal-level Proficient status must be relentlessly stressed because, in fact, the Nation is not apt to see immediate increases in the proportions of children attaining the Proficient or Advanced performance. NAGB must explain the honest fact that they *deliberately* set high performance demands represented by the Proficient and Advanced levels that *are not easily reached*—and won't be attained by most students for some time to come. In short, the public-relations task here is to help policymakers, and the publics they represent, understand more accurately just how challenging is the level of performance represented in what many citizens now regard as a somewhat easily attained level of quality.

The second communication tactic, to be initiated simultaneously, is to emphasize the need for students to be steadily moving toward the Proficient level. This can be done by having students make meaningful advances along the six "needs improvement" subcategories below the Proficient level. Great attention should be lavished on this new six-step improvement continuum, and great applause should be lavished on educational programs that can show progress according

to this more fine-grained improvement ladder. In addition, NAGB ought to do whatever it can, short of creating the image of a natural curriculum, to support educators who are trying to move more students toward proficiency. It *cannot* be assumed that the mere creation of six new reporting subcategories will automatically yield improved scores over time, either nationally or at the State level. Thus, so that there can be more celebrations of student *toward-goal* progress over time, NAGB needs to start thinking *instructionally* so that it can help those educators who wish to move children higher on NAEP's six-rung improvement ladder.

I concede that, in large measure, both of these communication tactics reside in the realm of public relations, but I believe that if the Board's members do not make a meaningful commitment to such communication activities, and both tactics are fundamentally only informational, then the current reporting-level dilemma faced by NAEP users will not be satisfactorily resolved. NAGB, admittedly with the best of intentions, will have dug itself into a dark, inescapable dungeon.

What is being recommended, to sum up, is a clarification-focused strategy that, because of a refinement in NAEP's below-goal reporting categories, will make it possible for NAEP reports to be a cause of celebration, not sorrow. If NAGB can mount an effective set of clarification activities, via print, video, the Internet, and any other suitable media, then this solution-strategy might just turn out to be one of those rare instances in which one can simultaneously have one's cake while gleefully consuming that same pastry.

No Regrets, No Illusions

I have appreciated the opportunity to bring a different perspective to NAGB's consideration of this important issue. But, of course, neophytes typically tend to shine less often than to stumble. I realize that some first-rate folks have dealt with this achievement levels issue over the years, and I have confidence that the Board's current members will draw on that experience to do what they think is best for the Nation's children. Nor do I have any illusions that my proposed solution-strategy will evoke uniform approval from all those concerned. It is, simply, the best solution I could come up with.

References

Alexander, L. and James, H. (1987). *The Nation's report card: Improving the assessment of student achievement.* Study Group on Updating the Nation's Report Card. Washington, DC.

Casserly, M. (1998). *Comments Regarding NAEP and Voluntary National Tests.* Presented at the meeting of the National Assessment Governing Board, Arlington, Virginia.

National Academy of Education (1988). *Commentary by a review committee.* Robert Glaser (Chairman), Washington, DC, p. 58.

National Assessment Governing Board (1989). *Staff paper on setting goals for the National Assessment.* Washington, DC.

National Assessment Governing Board (1990a). *Bulletin.* Washington, DC.

National Assessment Governing Board (1990b). Technical Methodology plus Analysis, 1990b Reporting and Dissemination Committees. *Setting achievement goals for the National Assessment of Educational Progress*. Washington, DC.

National Assessment Governing Board (1990c). *Setting appropriate achievement levels for the National Assessment of Educational Progress.* Washington, DC.

National Assessment Governing Board (1993). *Policy statement.* Washington, DC.

National Assessment Governing Board (1995). *Policy statement.* Washington, DC.

National Assessment Governing Board (1999). *Public hearings and written testimony on the purpose, intended use, definition of the term "Voluntary," and reporting of the proposed Voluntary National Test.* Washington, DC.

National Assessment Governing Board (2000). *Report on public perception of the National Assessment of Educational Progress achievement levels.* Washington, DC.

Nellhaus, J. (2000). "States with NAEP-like performance standards." In Bourque, M.L. *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements.* Washington, DC: National Assessment Governing Board.

SECTION 8

What NAEP's Publics Have To Say

Claudia Simmons and Munira Mwalimu

Aspen Systems Corporation Rockville, MD

November 2000

What NAEP's Publics Have To Say[1]

Claudia Simmons and Munira Mwalimu
Aspen Systems Corporation

Board Plan

The National Assessment Governing Board's (NAGB's) Report, sent to Congress on June 30, 1999, contained a provision for gathering information about the public's perception of the National Assessment of Educational Progress (NAEP) achievement levels. Because the public was identified as the primary audience from whom information would be gathered in the NAEP redesign plan, it was important to have a systematic approach for gathering such information about the three legislated criteria: whether the achievement levels are reasonable, valid, and informative to the public. Although evidence suggests that the levels are useful and informative, as acknowledged in the National Academy of Sciences' report (1998), the Board believed it would be helpful to hold several information-gathering sessions around the country with specific NAEP audiences.

Executive Summary

On behalf of the NAGB, Aspen Systems Corporation, a contractor to the Board, convened four discussion groups that were designed to gather information from targeted audiences on two of the three legislated criteria for achievement levels established by the Board:

1. Are achievement levels reasonable?
2. Are achievement levels informative?

The four categories of audiences identified by NAGB for the discussion groups were:

1. Governors' and States' legislative staff.
2. State assessment personnel.
3. Public and private educators, administrators, and parents.
4. Business leaders and education policymakers.

The first of the four discussion groups took place on September 9, 1999, in Atlanta, Georgia, in conjunction with a regularly scheduled meeting of the Southern Regional Education Board. The second discussion group was held on October 5, 1999, in Alexandria, Virginia, in conjunction with the regularly scheduled meeting of the Education Information Advisory Council. The third session was convened on November 22, 1999, at the San Francisco Unified School District Offices. The last session met on December 9, 1999, at the offices of the Greater Houston Partnership in Houston, Texas.

[1] Discussion groups convened by Aspen Systems Corporation under contract with the National Assessment Governing Board. This report was prepared by Claudia Simmons and Munira Mwalimu and does not reflect viewpoints of the National Assessment Governing Board.

Each discussion group was composed of no more than 12 participants and each session was scheduled to last no more than 2 hours. Recruitment of participants for the discussion groups was through convenience sampling, because the requirements for the audience types were targeted. NAGB also suggested convening the discussion groups when possible with other meetings or conferences to encourage and facilitate audience participation.

A packet providing background information and describing the purpose of the discussion was mailed to each participant in advance. Materials included an introductory letter that stated the purpose of the discussion group and provided information on the location, date, and time. Materials also included policy definitions of achievement levels and content descriptions in selected subject areas assessed by NAEP, such as mathematics, reading, writing, and science, so that participants would come prepared for the discussions.

All four discussion groups focused on the following two topics:

- The reasonableness of the NAEP achievement levels with regard to three components: (1) policy definitions of achievement levels, (2) content descriptions of achievement levels, (3) relation of NAEP achievement levels to other assessments.

- Audience experience and reaction to achievement levels, with the comments focusing on two areas: (1) reporting of achievement levels, and (2) usefulness of achievement levels.

The following summary highlights the findings from each category of participants for the two areas of discussion. This summary is followed by an in-depth report of the discussions that took place among each of the four groups. Feedback received from participants has not been attributed to any one participant or State to protect participants' confidentiality.

Discussion Topic 1. Reasonableness of Achievement Levels

Each discussion group was asked whether in the group's opinion, the achievement levels are reasonable with regard to the policy definitions, content descriptions, and the relation of the achievement levels to other assessments.

1. Policy Definitions of Achievement Levels

With regard to the policy definitions of achievement levels, staff representing Governors and State legislators who met in Atlanta believed the policy definitions were reasonable expectations for students. They believed the NAEP definition of Proficient represented above-grade performance and should be each student's goal. However, two participants in the Atlanta group believed a fourth level of Below Basic performance should be established to correlate more closely with the various States' definitions of student performance.

State assessment personnel attending the Alexandria discussion group expressed the viewpoint that what is reasonable is subjective and varies from individual to individual. For example, not everyone can agree that a particular skill is basic. In general, participants believed that the achievement levels represent goals or ideals rather than reflect actual

student performance. NAEP data, they said, indicated that few students are reaching the goals.

Educators, administrators, and parents who met in San Francisco concurred with State assessment personnel with regard to "reasonable" being a subjective matter. In addition, there appeared to be consensus on the need for NAEP to define a category of Below Basic, because a large number of students would fit in such a category. These participants also suggested renaming the Below Basic category to reflect student performance in a more positive manner. There was consensus among the San Francisco participants that the NAEP achievement levels represent very high standards, and that they go beyond the commonsense definitions of student performance at each level, particularly the Proficient level. However, the group called the high standards laudable.

Business leaders and education policymakers who met in Houston applauded the high standards set by NAEP through the achievement levels. They agreed with the definition of Basic as representing partial mastery of subject matter. This group believed the standards should not be lowered, and that schools, administrators, and teachers should help students attain the standards.

2. Content Descriptions of Achievement Levels

Participants were queried on the reasonableness of the content descriptions of the achievement levels.

The Atlanta group believed the content descriptions matched the policy statements and that they were reasonable and representative of what students at each of the three grade levels should know.

In contrast, the State assessment personnel raised concerns about the content descriptions, stating that the descriptions are value laden and do not accurately reflect the levels of student performance. They asserted that examples of items that were described as Basic were not actually basic, and exemplars of Advanced items were common-sense items. Some of the Alexandria participants suggested that the content descriptions themselves are not controversial; instead, it is the translation of these descriptions in percentages of students capable of performing at the level and their translation into curriculum descriptions that create problems. This group emphasized the importance of allowing standard setters or even the Board to modify the content descriptions as some States do to ensure their appropriateness. In addition, this group believed problems would arise in defining the content descriptions, such as using the word "proficient" in the meaning of Basic, creating more confusion.

The Atlanta group believed the descriptions set unrealistically high expectations for student performance, resulting in fewer students meeting the Advanced level of performance. This group also discussed the standard-setting procedures in their own States and how the States avoid confusion caused by definitions and terminology by identifying goals separately from defining standards. The assessment personnel also suggested that the composition of the standard-setting panels be reviewed to reflect more realistic definitions than those that appear

to be set by experts. In arriving at content descriptions, participants said consequences data should be provided to panelists to allow them to review their ratings prior to finalization. The State assessment personnel also remarked that NAGB sets achievement levels subject by subject, resulting in the lack of a holistic group viewpoint on the levels. They suggested that NAGB "design down," although they did not know how this could be done without going normative.

Discussants from the San Francisco group also said the exemplar items do not accurately reflect the level of performance defined in the content descriptions. This group raised concern over the value of the content descriptions because of the variance in curriculum taught in content areas throughout the country. Students' success on standardized tests depends on the curriculum they are taught. Curriculum decisions are made at local and State levels, and given this diversity, content descriptions at each grade level cannot be defined precisely.

Business leaders, policymakers, and education activists who met in Houston applauded the content descriptions, and recommended that the standards should not be lowered. They remarked that higher expectations of students would result in higher standards of performance. In the United States, they noted, there is an inclination to lower standards, whereas in many foreign countries, students are expected to meet higher standards.

3. Relation of NAEP Achievement Levels to Other Assessments

All four discussion groups agreed that the NAEP achievement levels cannot be compared with results from other standardized assessments, such as Advanced Placement (AP) tests, the Scholastic Assessment Test (SAT), or the American College Test (ACT). Each group gave the same reasons for this viewpoint. NAEP sampling methodologies are different from other assessments. Results from the other standardized tests are based on self-sampling, with students signing on individually and paying to take the test, whereas NAEP's sample is complicated and statistically defined. Further, the motivation for taking NAEP is different. The other tests are taken to secure college admission or earn college credit, whereas NAEP is taken simply to assess how students are doing.

Discussion Topic 2. Results: Are the Achievement Levels Informative to the Public?

The second topic participants discussed was their experience and reaction to the achievement levels with regard to two areas:

1. Reporting of achievement levels.
2. Usefulness of achievement levels.

A summary of each groups' viewpoints is provided below.

1. Reporting of Achievement Levels

The discussion on how achievement levels are reported in the individual States and cities represented by participants varied, although general similarities in reporting were apparent.

The Atlanta group reported that in recent years, there has been increased coverage of student performance on NAEP, reported through the achievement levels. This is attributed to the demand for more information on student performance. However, the larger State newspapers report the data rather than local newspapers. Local communities are more interested in data gathered and reported for the district and State assessments.

State assessment personnel who met in Alexandria made a significant number of comments on reporting achievement levels. In contrast to the Atlanta group, there was strong consensus on the need for NAEP data and the importance of the data to the States. However, the way the NAEP results are reported by the media raises a number of concerns:

- Since NAEP began reporting through achievement levels, NAEP is no longer viewed by the States as a "Dow Jones-like" barometer of student performance, because the NAEP achievement levels also represent future expectations of student performance. Further participating States have expectations of where they want their levels to be.

- The media often misinterpret NAEP data and report student performance negatively, even when the results are positive. The Federal Government can rectify this situation by taking steps to focus also on the positive results.

- Although achievement levels appear easy to understand and report, greater efforts need to be placed on understanding the data so they are accurately interpreted and reported through the media.

- States need assistance in interpreting and reporting NAEP data, perhaps through the provision of additional descriptive data and using the State Education Information Advisory Committee task force for input before the press conferences release results.

- NAEP results do not reflect variances in State processes such as differing educational practices, variations in curriculum, or student backgrounds. Therefore the results must be considered in context and trend data should be utilized.

- NAEP results do not show movement among the levels since the achievement levels are reported as scale scores. Descriptions can be supplemented with contextual information, a short statement on item mapping, and the continued release of sample items to help clarify scores.

- Achievement level reporting should include both scale scores and percentages. Scale scores are better at showing improvement over time and facilitating cross-State comparisons, whereas percentages are easier to understand.

Members of the San Francisco group—educators, administrators, and parents—were unanimous in their viewpoint that the media report little or no NAEP data. The few participants aware of NAEP reporting remarked that the media focused on negative student performance, and that this is sometimes attributed to the poor education offered by public schools. This group contended that the public schools are doing a good job. A similar viewpoint that this group shared with the State assessment group is that NAEP results do not reflect variances in State processes for establishing curriculum and fail to recognize different student backgrounds. The San Francisco group also applauded the release of NAEP items because it provides a clearer picture of what is being assessed and enables the public to understand the degree of difficulty for each item.

Business leaders and policymakers who met in Houston agreed that achievement levels should be reported through both percentages, which are easy for the public to understand, and scale scores, which show improvement over time and facilitate cross-State comparisons.

2. Usefulness of Achievement Levels

The final topic of discussion focused on the usefulness of the achievement levels as experienced by the participants of each discussion group.

There was a strong consensus among all four groups on the following points:

- NAEP achievement levels provide a common-sense approach to interpreting test results and offer a simplified explanation of student achievement.
- Achievement levels are easy to understand and interpret.
- Exemplars are useful in understanding the difficulty level of the NAEP assessment.
- Trend data collected by NAEP are useful.
- NAEP data are used to validate State standards.
- NAEP data should continue to be reported in percentages to help the public understand student performance.
- Tracking of different subgroups using the levels is helpful.
- In general, NAEP results are not used to leverage funding decisions.
- NAEP data are used by different audiences for different reasons. These audiences include policymakers, State boards of education, the media, parents, and educators.

Some States reported using NAEP as a model in developing NAEP-like test items used in the State and district assessments and in defining levels of student performance. The Atlanta and Alexandria groups felt particularly strongly that NAEP achievement levels help the States

understand the performance of students not only on the NAEP assessment but also in comparison to the states' performance compared with other States. These two groups also believed the NAEP results are used to influence State allocation of funding through the legislative bodies.

Participants at each discussion group were given the opportunity to make additional comments at the conclusion of each discussion. Comments ranged from NAEP matters in general, to the NAEP assessment methodology, achievement-level setting procedures, and the achievement levels themselves. These comments are included at the conclusion of the individual reports, which follow.

Report of Governors' and States' Legislative Staff

Westin Airport Hotel
Atlanta, Georgia
September 9, 1999

Participant Sampling

The first of the four discussion groups was convened on September 9, 1999, in advance of a regularly scheduled meeting of the Southern Regional Education Board (SREB) in Atlanta, Georgia. Recruitment of participants for the discussion group was conducted through a mailing prepared by SREB staff that included an announcement of the group. That mailing was included with others in preparation for the SREB meeting. SREB staff made follow-up telephone calls to recruit participants. Ten participants agreed to join the discussion group. Attachment A to this report provides a complete listing of participants. States represented were Arkansas, Georgia, Kentucky, North Carolina, Tennessee, Texas, and West Virginia.

The purpose of the Atlanta discussion group was to gather information on the perception of Governors' and legislative staff members on achievement levels, with regard to the reasonableness, usefulness, and utility of the achievement levels. Participants' comments have not been identified by name or by State to protect participants' confidentiality.

Results: Are Levels Reasonable?

The first area of discussion pertaining to the reasonableness of the levels focused on the following three points:

1. Policy definitions of achievement levels.
2. Content descriptions of achievement levels.
3. Relation of NAEP achievement levels to other assessments.

1. Policy Definitions of Achievement Levels

Participants noted that the policy definitions of the achievement levels were reasonable expectations for students. There was much discussion about various State-level definitions of student competence that were similar to the NAEP levels. Discussion participants agreed that the NAEP Proficient level should be the goal for all students. Most participants identified a similar State level of competence for students in their States. For example, two participants from different States identified four levels of competence for students. These four levels included an undefined NAEP level, which is referred to as Below Basic.

State A	State B	NAEP Level
Novice	D/F	Below Basic (not a specified NAEP level)
Apprentice	C	Basic
Proficient	B	Proficient
Distinguished	A	Advanced

Participants expressed a desire for a crosswalk between the NAEP levels and grade-level expectations by State. They noted that standards-based education at the local level was more focused on grade-level competency. Participants expressed the view that NAEP's Proficient level was often above grade-level mastery of material.

Participants emphasized the importance of terminology and said there should be a correlation between grade-level mastery and the NAEP achievement levels. However, the point was raised that there should be a Below Basic level of achievement defined to correlate more closely with State measures of student performance.

2. Content Descriptions of Achievement Levels

The Governors' and States' legislative staff found the content descriptions to be reasonable and representative of what students at each grade level should know. They believed the content descriptions did match the policy statements. Using the achievement levels as the guide, Basic, Proficient, and Advanced do specify the levels of achievement as defined in the policy statements. Concerns were voiced that the NAEP scores only reflect State performance, yet there was intense interest (at the State level) in what was happening in schools at the district level, perhaps even at the building level.

The difference between NAEP and State assessments lies in consequences. Whereas there are no consequences of student performance on NAEP, State-level reform efforts are based in large part on consequences. For example, if a school/district does not perform to *X* standard, then *Y* will happen as a consequence (e.g., the district will be taken over by the State or the teaching staff at the school will be reconstituted). The NAEP scores are reported, but there is no consequence for how the various districts perform because the results are aggregated by State. Districts can use test scores on other measures to leverage policy changes using the same rationale. If NAEP scores were reported at the district level, their importance would be different.

Participants liked the fact that real test items are released by NAEP. They believed the test items, more so than scores, had resonance with parents. Parents could identify with specific items. One participant mentioned that the NAEP scores could also be used to leverage more funding for higher education, because more parents may use opportunities to advance their own education as a way to better understand reporting of the scores.

3. Relation of NAEP Achievement Levels to Other Assessments

Participants found it difficult to compare the NAEP achievement levels with scores of other assessments such as AP scores, SAT scores, ACT scores, or scores on State/local assessments. The discussion group believed these assessments were not aligned with the NAEP assessment. Discussants reported that some policymakers question NAEP when comparing NAEP to other assessments. Participants identified several underlying issues:

- A lack of public understanding of different forms of assessment.

- A lack of public understanding of standardized test scores such as those provided through the SAT and ACT assessments. These assessments evaluate a self-selected sample and thus should not be compared with NAEP scores.

- A distrust of criterion referenced tests (CRTs) on some level by policymakers.

There was group consensus that NAEP scores cannot and should not be compared to AP, SAT, or ACT scores. Discussion participants found the NAEP expectations of student performance to be practical and meaningful.

Results: Are Levels Informative to the Public?

The second area of discussion focused on audience experience and reaction with regard to the achievement levels. Comments focused on two main areas:

1. Reporting of achievement levels in the States.
2. Usefulness of achievement levels.

1. Reporting of Achievement Levels

Participants pointed out that the media in their States do not give as much coverage to NAEP data as they do to district/State assessment data. The larger state newspapers are the ones that cover NAEP. Those in smaller communities without easy access to larger papers may miss the coverage. The public's response is influenced primarily by press coverage. Response by the public is usually favorable if the trend data are positive for the State.

In recent years, the levels are receiving more press coverage and have become of more interest to the public. Participants mentioned that reporting of the levels by the press was variable—sometimes more and sometimes less coverage. Again, since the results are reported at the State level as opposed to the district or school level, the reporting focus is slightly different.

2. Usefulness of Achievement Levels

Participants reported several uses for the achievement levels in the States they represented. All States represented reported student performance using the NAEP results. However, discussion members believed the impact of the NAEP results was somewhat limited because the results are not reported at the district level. Here, too, respondents acknowledged that results at the district level often have more resonance with a larger audience.

The following points highlight the usefulness and impact of the achievement levels in the States represented by participants:

- NAEP achievement levels are used to validate State assessment standards.

- States find the trend data very useful. Looking at longitudinal results by States is helpful in developing and refining educational policy in States, and the results are used as a "driver" to add impetus to State reform education efforts.

- States have used NAEP as a model in developing NAEP-like test items that have been used in high school-level district assessments.

- The NAEP achievement levels help States understand the performance of students on the NAEP assessments as well as the State assessments.

- All States represented indicated that the tracking of different subgroups using the levels was helpful for State/district planning. However, the States do not use the NAEP data in this way, although there is value in disaggregated data being reported.

- Discussion group members were of the viewpoint that NAEP results can be used to influence State resource allocation. Of particular interest was a comment about the influence on funding for higher education. For example, NAEP results at the 12th-grade level may indicate the need for more remedial courses.

- The levels are used with policymakers, State boards of education, the press, parents, and higher education programs in assisting with program development.

- NAEP data fueled State board discussions.

- Understanding of NAEP has not changed as a result of the levels.

Additional Comments

Participants made a variety of additional comments, as noted below:

- Participants expressed the need for further education of various audiences within the States to help them better understand the NAEP data.

- If resources allowed, the States would like more interpretation of NAEP results. NAEP data would have more uses if the data were manipulated in more ways. They would like for NAGB to do more analysis to assist the States in their interpretation of results.

- States would like more information on the validity of the measures. Questions arise on interrater reliability that they would like addressed.

- Term limits influence what happens with policymakers. There is more turnover in State legislatures because of term limits, so policymakers have to be educated about NAEP more frequently.

Attachment A

Discussion Group Participants: Atlanta, GA
September 9, 1999

Brad Borum
Policy Analyst
Senate Research Office
204 Legislative Office Building
18 Capitol Square
Atlanta, GA 30334
Phone: 404–656–0015
Fax: 404–657–0929
Bborum@inet.legis.state.ga.us

Audrey Carr
Legislative Analyst
Legislative Research Commission
Annex Building, Room 105
State Capitol
Frankfort, KY 40601
Phone: 502–564–8100
Fax: 502–564–6543
Audreycarr@lrc.state.ky.us

Jack Elrod
Committee Director
Texas Senate Education Committee
Sam Houston Building, Room 440
201 East 14th Street
Austin, TX 78702
Phone: 512–463–0355
Fax: 512–463–7567
Jack elrod_sc@senate.state.tx.us

Sammy Gray
Legislative Analyst
West Virginia Senate
Building 1, Room 427M
1900 Kanawha Boulevard, East
Charleston, WV 25305
Phone: 304–357–7952
Fax: 304–357–7881
Graysa@mail.wvnet.edu

Hank Hager
Attorney
Senate Education Committee
State Capitol, Room 415 M
1900 Kanawha Boulevard, East
Charleston, WV 25305
Phone: 304–357–7871
Fax: 304–357–7881

Connie Hardin
Director
Office of Legislative Budget Analysis
G–9 War Memorial Building
Nashville, TN 37243
Phone: 615–741–4378
Fax: 615–253–0189
Connie.hardin@legislature.state.tn.us

Mark Hudson
Legislative Analyst
Arkansas Bureau of Legislative Research
State Capitol, Room 315
Little Rock, AR 72201
Phone: 501–682–1937
Fax: 501–682–9626
Mark@arkleg.state.ar.us

Jim Johnson
Senior Fiscal Analyst
North Carolina General Assembly
Legislative Office Building, Room 619
Raleigh, NC 27603
Phone: 919–733–4910
Fax: 919–715–3589
Jimj@ms.ncga.state.nc.us

Michael Murdoch
Senior Policy Analyst
206 Legislative Office Building
18 Capitol Square
Atlanta, GA 30334
Phone: 404–657–4604
Fax: 404–657–4606
Mmurdoch@legis.state.ga.us

Jacqueline Nash
Research Analyst
Senate Education Committee
Tennessee General Assembly
Nashville, TN 37205
Phone: 615–741–3038
Fax: 615–741–9349
9A Legislative Plaza
Jacqueline.nash@legislature.state.tn.us

Report of State Assessment Directors and Staff

Holiday Inn Select Old Town
Alexandria, Virginia
October 5, 1999

Participant Sampling

The second of the four discussion groups was convened on October 5, 1999, in conjunction with a regularly scheduled meeting of the Education Information Advisory Council (EIAC) meeting in Alexandria, Virginia. Recruitment of participants for the discussion group was conducted through a mailing to EIAC members scheduled to attend the meeting. Participants had the opportunity to register for the discussion group via e-mail to Aspen staff. Follow-up telephone calls to recruit participants were made by Aspen staff. In-person recruitment of participants took place at the EIAC meeting via NAGB staff onsite at the meeting. The discussion group had eight participants. Attachment B to this report provides a complete listing of participants. States represented were California, Connecticut, Idaho, North Carolina, Ohio, Texas, Washington, and West Virginia.

The purpose of the Alexandria discussion group was to gather information from State assessment directors and staff on the reasonableness, usefulness, and utility of achievement levels. Participants' comments have not been identified by name or by State to protect participants' confidentiality.

Results: Are Levels Reasonable?

The first area of discussion pertaining to the reasonableness of the levels focused on the following three points:

1. Policy definitions of achievement levels.
2. Content descriptions of achievement levels.
3. Relation of NAEP achievement levels to other assessments.

1. Policy Definitions of Achievement Levels

In discussing the reasonableness of the achievement levels, several participants remarked that reasonableness is a perception issue and depends on each individual's understanding of what is reasonable. Nearly everyone has an individual opinion about what is reasonable. One participant noted that the definition of Basic as partial mastery involves making decisions regarding components of mastery, and these are individual perceptions. Another participant remarked that while one can read the definition and understand what it means, when it is explained to someone, that person might not agree with the definition of Basic. Several participants alleged that the media, for example, generally interpret proficiency as basic performance or some level of competence.

One participant noted that there is a difference between needs and wants as opposed to the ideal. What people desire is different from what they need, and NAEP confuses the two, he said.

What goals will be needed for American education in the next century are only hypothesis at this point. However, to help students reach these new goals, NAEP has adopted high student performance standards for the various subjects measured by the NAEP assessment. Current NAEP data show that few students are reaching these goals.

2. Content Descriptions of Achievement Levels

One participant stated that the descriptions are acceptable, but the translations of the descriptions into percentages of students who can do a certain level of work create problems.

Many participants believed the words "basic" and "proficient" get in the way and create problems. For example, it is not true that 75% of students cannot read. The words are value laden. The question is, what is better?

One participant remarked that it is not the descriptions or the standard-setting method that lead to high standards. Rather, it is the translations of Proficient descriptions into curriculum subject-matter descriptions that create confusion. "If you look at the science descriptions that say what Advanced means and what Proficient means, and if you read just the eighth-grade Advanced or Proficient descriptions . . . you'd think this student has studied science in college, and we're talking about eighth grade students," the participant remarked. "And then you get standard setters sitting down and looking at that, and that's what they use when they start making decisions about items, and test content, and where to place these levels. They're using the curriculum descriptives and I don't think it's this that's causing the problem." Another participant remarked that it is a combination of both.

An additional viewpoint expressed was that until the science assessment, the standard setters were allowed to change the descriptions, so that what was used to create the tests was not necessarily the final descriptions used to set the standards.

One participant observed that the Basic level of performance is set very high. He illustrated this point by referencing the content description of Basic at grade 4 on page 59 of the Science Performance Standards report. He noted that the descriptions in the two white paragraphs following the gray paragraph are at a very high level and represent an Advanced level of understanding, yet this is supposed to reflect the Basic level of performance. He remarked that the reason there are so few percentages of students at the Advanced level is because of the way the content descriptions are written, not just in science, but in other subjects as well.

Another participant added that people blame the achievement levels on the standard-setting methodology when the descriptions themselves set very high expectations for student performance. Some participants believed that the subject areas are a moving target and that there is a problem with test development that has many standards but no movement. For example, science is not a well-defined subject and has a broad domain. It is not well defined in terms of specific chunks of knowledge, in a whole context from one test administration to another. As a result each administration has many different items from the prior assessment. For example, when Educational Testing Service (ETS) designs the NAEP Science Assessment, some items are rejected after the field tests. After ETS makes changes to the items, they are put back in the

regular administration, yet the statistical values of these changed items are unknown. ETS then scales everything, presumably to resolve the problems, but the problems do not necessarily go away.

Participants identified other problems as well. One participant remarked that the exemplar on the life cycle of the butterfly on page 9 of the Science report defines this as a Basic item—drawing and labeling the missing part of the picture that is the cocoon. This is a Basic item yet it is clearly not Basic because it requires a more thorough understanding of the life cycle of a butterfly. He noted that the exemplar on page 10 of the Proficient item is actually a far easier item than the one labeled as Basic. On the other hand, the exemplar on page 11 is just a recall, common-sense item, without student need for scientific knowledge.

Participants believed the development of the test therefore has serious technical issues, which complicate the achievement levels, leading the National Academy of Sciences to state that the achievement levels are "fundamentally flawed." This goes back to the test design and descriptions, yet the standard-setting methodology is blamed.

One participant noted that in his State, four standard-setting committees in four different areas set the standards, and there were various levels. Standard setters were told to ensure appropriateness across content areas. Once the information was provided to the department, recommendations were raised and lowered so that everything was appropriate. Data were also looked at across the grades to have a roller coaster effect.

When the curriculum design is written and tests are developed based on the frameworks, then standard-setting committees should be allowed to change the descriptions because there may be confusion along the way, participants asserted. Overlaying content creates a conjunctive effect that continues through to standard setting.

Participants said that once the standard setters make the recommendations, the Board needs to review them again for appropriateness, as it did with the science results. The participants contended that for science, they were lowered for political reasons, but that this needs to be done consistently across all content areas. With the Board having a policymaking role, some participants were doubtful that the Board would want to be viewed as lowering standards. The Board also states that the levels are experimental (in its report to the National Academy of Sciences), yet no one working in the field views them as experimental.

An additional problem with the content descriptions raised by one participant was that in the definition of the Basic level, the word proficient is in the middle of the definition, which is confusing. The Proficient level is considered competent, so this is confusing. It is clear that for the Advanced level, superior is superior, but if a layman was asked to rank the performance of students as competent and proficient, it is hard to tell which of the two descriptors is superior. If one is Proficient, then one is competent, but if one is Basic, one is nearly proficient, which is partial mastery. One participant remarked that educators have ruined the content descriptions.

Descriptors need to provide a picture of the project goal and image of student performance across grade levels. The curriculum standards are set in isolation. Participants discussed whether the levels need to be anchored to reality, or whether it is not reality but an ideal. The ideal can be a very high standard, which is why there may be a conflict between words, leading some to conclude that that is why we are not reaching the ideal.

The composition of the panels setting the standards also needs to be examined. In all content areas, the standards are too high; this all comes back to the descriptions, and is blamed on the standard-setting methodology. The public in the panels defers to the experts. After the panels define the content, the descriptions need to be read. Teachers and people who know the developmental level of thinking for students would be ideally suited to do this.

All standard-setting methods rely on judgments and composition of groups. The reality of schools is not portrayed. Subjects are not covered in isolation, and students have to learn other things, too. It is unlikely that students will accomplish complete mastery of any subject matter. However, the subject matter experts appear to assume that students will gain total mastery of various content areas.

In one State, recommendations made by the standard-setting committees were made at various levels, across content areas and across grades. The State took the recommendations and made decisions to level them out; someone needs to look at achievement levels across content areas as well.

Participants remarked that confusion in the definitions and terminology are avoided in the States by stating a goal or standard, then defining the various levels of performance, such as Levels 1, 2, 3, and 4. In one State, levels are tied to the purposes of the test per grade level. The only verbal description is that a student meets the standard or is below the standard. Another State defines student performance via different levels of mastery, such as partial, inconsistent, and consistent mastery, with one level being defined as the ultimate goal.

Participants also talked about providing consequences data to participants. In at least two prior NAEP assessments, consequences data were provided to the panelists after they had made their final decisions, then they were allowed to say whether they would have changed their ratings if they could. The vast majority said they would not have changed the ratings. For the past assessment, they were for the first time given consequences data before they made their final decisions. All the participants believed this should always be the case, so that panelists can review their ratings before they are finalized as it provides a reality check.

Further, because NAGB sets achievement levels subject by subject, there is no whole group viewpoint. Participants believed there is a need to "design down," but the issue is how to do this without going normative.

3. Relation of NAEP Achievement Levels to Other Assessments

In response to the question on how NAEP relates to other assessments such as the ACT, SAT, or AP tests, participants stated that the other tests cannot be equated to NAEP, since the stakes are

different. The other tests are self-selecting tests that students pay to take, and the motivation is often different as well. The other tests are taken to secure admission to college and the AP tests are taken to get college credit. The settings are therefore different.

Participants also remarked that SAT or AP tests cannot be used to validate NAEP data since the motivations for taking the tests are different.

Results: Are Levels Informative to the Public?

The second area of discussion focused on audience experience and reaction with regard to the achievement levels. Comments focused on two main areas:

1. Reporting of achievement levels.
2. Usefulness of achievement levels.

1. Reporting of Achievement Levels

There was unanimous consensus among participants that the media misinterpret the results. One participant noted that when NAEP was set up to be the Nation's Report Card, it was comparable to the Dow Jones. Performance is to be articulated through student exemplars but once performance levels come in, these are future expectations of where we want to be, so performance cannot be compared to something like the Dow Jones that reports on current stock market activities.

There was strong consensus within the group that reporting on NAEP results was almost always negative. Bleak pictures always appeared to be painted on student performance, even though student performance in some data releases shows significant improvement. Participants believed the Federal Government puts a negative spin on the results for their own purposes, perhaps to seek additional funding for education.

Achievement levels are easy to misinterpret. They appear to be very easy to understand, but many steps go into the process of understanding achievement levels. There is a greater burden on the media to provide accurate information so that it is not misinterpreted. For example, the recent writing assessment release indicated that the vast majority of students can write at the Basic or higher level, yet the message in the press release from the National Center for Education Statistics (NCES) was negative, taking the approach of "the half-empty glass as opposed to the half-full glass."

With regard to the writing results, NCES had a very negative spin. One participant reported that the State he represented wrote its own more positive press release. Only one newspaper reported the writing results negatively, but the rest of the State newspapers picked up the positive results. The data can be used positively. Perhaps the fact that "good news does not sell" is what drives negative reporting of results. Participants noted that the Federal Government appears to want negative news, as it would be hard to push for more funding if students are doing well. Persons in politics get the credit for good news. In one State the results were widely reported even though no press release was prepared, evidencing the interest of the public in the results. Writing is, however, not considered as hot a topic as mathematics or science, so the impact of the writing release was not as significant to the public.

One participant continued to be troubled by the message that most kids fail and very few kids are at the Proficient level. Another participant stated that this did not trouble him, since he views the achievement levels as a moving target and looks at the standards as the goal. Another remarked that the spin put on the numbers is bothersome. Unreasonable goals are set and they portray the public schools as failing, which is not true.

Participants urged that the press releases written by the Federal Government convey positive information rather than negative, especially where the data are indeed mostly positive. If the States were provided with more descriptive data, they might help in the interpretation of results, perhaps by holding a prerelease briefing. Aspen staff reminded participants that NCES had recently provided a briefing prior to release of the writing results; a media person and a State assessment person were invited from each State. Participants stated that this briefing was held too close to the timing of the press conference. Thus it appeared that the briefing was to simply inform the States on what was being done, rather than to seek input. One participant noted that in his experience, the wording of the press release was changed slightly to accommodate his State's request, so the input process was useful for his State. Participants believed it was necessary to let the States know the Federal Government's spin on the results. An alternative suggestion made by one participant was to use EIAC task force members for input in advance of each press conference by providing them with embargoed information so that they can provide comments then bring them to the group.

Further, the results do not show variances in State processes. For example, the education system in State A could be superior to the one in State B, yet the performance of students in State B is always in the top range. Different educational practices could be undertaken in one State as opposed to another. Further, the student backgrounds are different, and this does not show up in the results. In addition, curriculums vary in each State.

In one State, the fact that about 30% of high school graduates attend 4-year colleges could be interpreted to indicate that students are in fact Proficient and that the proficiency level is when students are primed to go into a 4-year college. In another State 47% of the students are passing the reading test. Therefore the results are not negative at all. Trend data make a difference; in isolation none of this makes sense.

Another concern expressed by the participants was that the three categories of achievement level scale scores could move, yet students could remain in one level for a long time and not show any improvement when results are released. There is therefore a difference between what is needed, what is acceptable, and what is ideal.

With regard to the scale score line, a short statement on item mapping can provide a richer description of what students can do at each level. There is a need to put "bones" on what the test is, there is a need for contextual information; descriptions alone are not enough, because they do not really convey the message. Exemplar items are good because they provide context. Participants suggested mapping sample items on the probability scale to provide one picture. One participant remarked that in his State, sample items were released after an assessment. Prior to the assessment, everyone said the sample items were very easy. Once they were released, their

viewpoints changed, as even people with master's degrees could not provide correct answers. Similarly, NAEP can dispel the myth about student performance by releasing the items.

Regarding the question of reporting by scale scores as opposed to percentages, participants said the question is flawed because the response depends on the context in which the levels are used. For public reporting, percentages are easier to understand, whereas scale scores are good for showing improvement over time. The importance is in the standard errors; if there is a significant change, the standard errors need to be examined to determine if real growth has occurred. Reporting by both scale scores and percentages is necessary because both can be useful, depending on the context. Scale scores facilitate cross-State comparisons, while percentages do not.

2. Usefulness of Achievement Levels

The following points highlight the usefulness and impact of the achievement levels in the States represented by the participants.

One participant reported that the performance standards are very useful in his State because they serve two purposes:

- Provide an indication of where students are.
- Provide information on how students should perform and what they should be able to do.

With regard to the overall usefulness of the achievement levels, there was consensus within the group that everyone pays attention to them. There was specific agreement on the following points:

- The achievement levels are assessable and provide a common-sense approach to interpreting results. Their appeal is the uncomplicated, simplified explanation of student achievement.

- They are easy to interpret and are used by both critics and proponents.

- Laypeople can better understand the achievement levels in terms of percentages as opposed to the NAEP scale. They can relate to whether student performance is near, at, above, or below the achievement level. People create their own categories of scores, and that is how they relate.

- The NAEP achievement levels are used as an outside validation of State assessment standards and for determining whether what the State is measuring is comparable to what NAEP is measuring. This is especially important where the State assessments are all criterion reference tests.

- The NAEP achievement levels help the States understand the performance of students on the NAEP assessments as well as on the state assessments.
- All States represented indicated that they tracked different subgroups using the levels, especially with regard to gender and race.

- The levels are used with policymakers, legislators, State boards of education, the press, parents, and State education staff, each for their own purposes. For example, parents may use them to determine how the State is doing compared with another State or the Nation as a whole, while legislators may use them to advocate for more funding for areas needing improvement. In one State, the achievement levels are used by legislators, such as the House Education Committee, and the Governor to substantiate their points of view to advocate more funding for charter schools and tutoring programs and to leverage other resources to improve the educational system.

- For the majority of participants, the achievement levels are not used to leverage funding decisions. In one State, however, a participant reported that the levels are used as a "club" by policymakers for a State voucher initiative. This is not a function of achievement levels, but rather a function of how the results are used.

- Achievement levels are used as a model in a few of the States represented. In one State example, the Proficient definition of performance is almost exactly the same as the NAEP definition of Proficient. In another State, the State standards are set based on the NAEP process of setting performance standards as Basic, Proficient, and Advanced, although there is no actual relationship and link in the process to the NAEP process. Some call it a modified Angoff method. Other States use the NAEP framework more than the achievement levels.

- NAEP provides information to the States on how each State ranks nationally.

- NAEP results are not used to determine policy other than as validation.

- Audiences for the achievement levels are all segments of the population.

- Parents are more interested in looking at the rankings than the achievement levels.

- Educators and teachers often ask for more detailed information such as the sample items. Other groups are satisfied with just the numbers; this is part of the problem with overlaying definitions of the achievement levels.

Additional Comments

Participants made a variety of additional comments, as noted below:

- There is a need for more balance between standards-based reporting and normative-based reporting that is more descriptive. Illustrative items might help dispel the notion that students are failing.

- There is a need to emphasize the positive where the data support it, rather than consistently showing poor performance of students even when the results indicate improvement.

- The test development process undertaken by the current NAEP contractor, ETS, is flawed. For example, the scoring guides for the open-ended items are assigned a numerical value of

0, 1, 2, or 3 or so on in a list fashion, instead of a good annotation to understand the context. It is the most simplistic scoring guide ever, and it is focused more on quantity than to quality.

- There are many problems with test construction, item development, pilot testing, and scaling of the assessment, yet the achievement levels are getting the blame and being termed "fundamentally flawed."

- The current achievement level-setting process seems to have a missing link. NAGB has responsibility for the content framework, NCES has responsibility for development and scoring (through a contractor), and NAGB sets achievement levels. Somewhere along the line, there needs to be greater coordination of the entire process. There is a need for constant feedback, as it seems the organizations may be working at cross-purposes. There are technical and communication problems. For example, ACT, the NAGB contractor that sets achievement levels, cannot contact ETS, the test developer, without going through NCES. Such communication procedures cannot facilitate good interaction, as they set up impossible procedures; perhaps this is a carryover from the achievement levels battle.

- Achievement levels are needed and useful to all States represented by the discussants.

Attachment B

Discussion Group Participants: Alexandria, VA
October 5, 1999

Cordelia Alexander
Director, Test Center
Department of Testing and Campus Support
Dallas Independent School District
3801 Herschel Avenue
Dallas, TX 75219
Phone: 972–749–2710
Fax: 214–357–4602
coralex@swbell.net

Kris Kaase
Consultant
Division of Accountability Services/
Reporting Section
NC Department of Public Instruction
301 North Wilmington Street
Raleigh, NC 27601–1203
Phone: 919–715–1203
Fax: 919–715–1204
kkaase@dpi.state.nc.us

Duncan MacQuarrie
Director, Assessment and Evaluation
Office of Superintendent of Public Instruction
P.O. Box 47200
600 Washington Street SE.
Old Capitol Building
Olympia, WA 98504–7200
Phone: 360–743–3449
Fax: 360–586–2728
duncanm@ospi.wednet.edu

Karen Nicholson
Assistant Director
West Virginia Department of Education
Building 6, Room D–057
1900 Kanawha Boulevard
Charleston, WV 25305
Phone: 304–558–2651
Fax: 304–558–1613
knichols@access.k12.wv.us

Douglas Rindone
Chief, Bureau of Student
Assessment and Research
Connecticut State Department of Education
P.O. Box 2219
165 Capitol Avenue
Hartford, CT 06106
Phone: 860–566–1684
Fax: 860–566–1625
douglas.rindone@po.state.ct.us

Gerry Shelton
Administrator
Standards and Assessment Division
California Department of Education
721 Capitol Mall, Sixth floor
Sacramento, CA 95814
Phone: 916–657–3011
Fax: 916–657–4964
gshelton@cde.ca.gov

Sally Tiel
Coordinator, Assessment and Evaluation
Idaho Department of Education
Len B. Jordan Building
P.O. Box 83720
650 West State Street
Boise, ID 83720–0027
Phone: 208–332–6943
Fax: 208–332–6965
srtiel@sde.state.id.us

Roger Trent
Director, Assessment and Evaluation
Ohio Department of Education
Room 207, 65 South Front Street
Columbus, OH 43215–4183
Phone: 614–466–3224
Fax: 614–728–7434
Ae_trent@ode.ohio.gov

Report of Administrators, Teachers, and Parents

San Francisco Unified School District Offices
San Francisco, California
November 22, 1999

Participant Sampling

NAGB held the third of its four discussion groups on November 22, 1999, in San Francisco, California. Recruitment of discussion group participants was conducted by the San Francisco Unified School District, through the joint efforts of a supervisor from the Assessment Office and the Associate Superintendent for Elementary Programs. Follow-up telephone calls to recruit the required participants were made by Aspen staff.

The discussion group had 12 participants, with 3 observers. Attachment C provides a complete list of participants. Three-member teams composed of an administrator, teacher, and parent represented the following four schools: San Francisco (an independent school, grades K–8), Junipera Sera Elementary School (grades K–5), Francisco Middle School (grades 6–8), and Lowell High School (grades 9–12).

The purpose of the San Francisco discussion group was to gather information on the perception of administrators, teachers, and parents on the reasonableness, usefulness, and utility of student achievement levels. Participants' comments have not been identified by name or by State to protect participants' confidentiality.

Results: Are Levels Reasonable?

The initial discussion concerning the reasonableness of achievement levels focused on three main points:

1. Policy definitions of achievement levels.
2. Content descriptions of achievement levels.
3. Relation of NAEP achievement levels to other assessments.

1. Policy Definitions of Achievement Levels

The discussion began with several participants observing that reasonableness is a perception issue that depends on an individual's understanding of the reasonableness issues. This led to the suggestion that there should be a category of Below Basic because of the large number of students who fit within that category. The issue then shifted to whether the NAEP definition of Basic was too high or too demanding. One participant, for example, expressed concern over the issue of credibility when discussing the science scores of 12th graders. He noted that, as currently defined, there could be a perception that all 12th graders in the country were Below Basic in science. The participants believed this was not a perception based on the reality of how students progress in science. The discussion then centered on the possibility of modifying the definition of Basic, or lowering the Basic cut-score, so that more students would meet the

standard. Because the proposed category would reflect the performance of below basic students, it was recommended that a more positive title be developed for this category.

Another participant observed that the description of Proficient, too, seemed to reflect a higher level of student performance than seemingly indicated by the common-sense definition of Proficient. Nevertheless, other participants pointed out that the definition of Basic indicates that there is at least a partial mastery of prerequisite skills at each grade level, and that students are asked to communicate increasingly sophisticated ideas as the levels progress from Basic to Proficient to Advanced.

One parent noted that a child's developmental level may influence the "demonstration of competency" at various levels. For example, a parent might wonder if his or her child scored Below Basic because the student was not developmentally ready for abstract thinking or because the school was failing to teach the appropriate material. This parent asserted that the levels should specifically address a child's ability to perceive abstract ideas and to interpret abstract concepts. Further, the participant noted that "just because a student is in eighth grade, he or she may not necessarily be an abstract thinker, while it could be generally assumed that all students in the 12th grade would be abstract thinkers."

Group members agreed that because NAEP has set very high performance standards, few students can measure up to expectations. Nevertheless, participants insisted that high standards are laudable goals.

2. Content Descriptions of Achievement Levels

The group's discussion focused next on the reasonableness of standards. One participant found the descriptions only somewhat helpful because of the variance in curriculum taught in various content areas throughout the country. Students' success on standardized assessment tests, therefore, depends on the type of material students are taught. Participants said it is the translations of descriptions into percentages of students who work at a particular level (Advanced, Proficient, and Basic) that are problematic.

The discussion group noted that there is variance throughout the country in the role that school districts and specific State departments of education play in determining local curriculums. In some instances, there is a significant amount of local control at the district level and all decisions about curriculum are made at that level. In other instances, some State departments of education dictate policies and procedures and specify the curriculums to the districts. Given this diversity in determining what is taught, participants believed it would be difficult to develop an assessment measure that can adequately address the many variables in curriculum for each district.

This concern arose from a discussion of what science elements are covered at what grade level. One participant noted that in the San Francisco Unified School District, the material on oceanography is not covered until seventh or eighth grade, yet it is on the NAEP science fourth-grade assessment. NAGB staff informed participants that the U.S. Department of Education opposes a national curriculum, although the Department recognizes that variations in curriculums at different grade levels can influence performance on various assessments.

Another participant noted that one of the science content descriptions defines grade 4 performance in science at a very high level, representing an Advanced level of understanding, yet the description is supposed to represent of the Basic level of performance. For example, the group questioned whether the exemplar on the life cycle of the butterfly on page 9 of the Science report is really a Basic item, as described. Participants believed this item requires a more thorough understanding of the life cycle than would be attained at the Basic level. Consequently, there may be technical concerns with test development. A group member suggested that perhaps so few students score at the Advanced level because of the way the content descriptions are written in all subject areas.

Other questions raised by participants included the following:

- Since there is no national curriculum, how were the assessments correlated to a sample of curriculums from throughout the country?

- Are non-English-speaking students administered the test in the language in which they are proficient?

- Are States required to participate in the NAEP assessments?

- How are the State samples selected?

- Can the cut-scores be adjusted to be more reflective of actual student performance?

In response to a question about report release dates, NAGB staff informed the group that the release date is usually a little more than a year after the assessment. Participants were advised by NAGB staff to access the NCES Web site at http://www.nces.ed.gov for additional information, such as the released test items. Participants who inquired about the rubric used to score the assessments were told that this information is also available via the NCES Web site.

3. Relation of NAEP Achievement Levels to Other Assessments

In a discussion about how NAEP relates to other assessments such as the ACT, SAT, or AP tests, participants agreed that it is difficult to correlate the other tests to NAEP, because students usually self-select the other assessments. These tests are usually taken to secure admission to college and the AP tests are taken to get college credit. However, there was consensus that the students who do well on these measures are usually the same students who do well on the NAEP assessment in the Proficient or Advanced ranges. For example, a score of 3 on a scale of 1–5 on an AP test would certainly indicate mastery of the material in a subject such as U.S. History, and perhaps be reflective of a grade of A in a regular U.S. History class.

Results: Are Levels Informative to the Public?

The second area of discussion focused on audience experience and reaction with regard to the achievement levels. Comments focused on two main areas:

1. Reporting of achievement levels in California.
2. Usefulness of achievement levels.

1. Reporting of Achievement Levels

There was unanimous consensus among participants that there is little coverage of the results by the media in the San Francisco area.

The group was concerned about the message portrayed by the media that most students fail and that very few students are at the Proficient level on these assessment measures. Participants discussed whether the NAEP assessment goals are unreasonable and lead to the public's false perception that the schools are failing. As a group, however, they believed strongly that the public schools were doing an adequate job in educating students. There was also discussion on the involvement of independent schools, but it was explained that it is difficult to evaluate independent schools because many do not have curriculums that match national standards. It was pointed out that parochial schools are included in the NAEP assessment.

Participants noted that the assessment results do not reflect either variances in State processes for establishing curriculums or the fact that students' backgrounds are often different. Participants had the impression that States in which the students had relatively homogeneous backgrounds, such as those in California tended to have higher performance.

Participants agreed that the release of the sample items was helpful because they provide a clear picture of what is being assessed and enhance the public's understanding of the degree of difficulty of the assessment items.

2. Usefulness of Achievement Levels

The discussion described in this section highlights the usefulness and impact of the achievement levels in schools represented by discussion group participants.

Participants reported that the performance standards are very useful and serve two purposes:

- Provide an indication of where students are.

- Provide information on how students should perform and at what level they should be.

There was consensus within the group that everyone would pay attention to the achievement levels if they knew what they were. It was noted that the material might be most relevant to the district staff in assessment offices.

Participants also made the following points on the usefulness of NAEP achievement levels:

- Achievement levels provide a common-sense approach to interpreting test results and offer a simplified explanation of student achievement.

- Achievement levels are easy to interpret with explanation, especially for the layperson.

- People outside the assessment field (laypeople) can understand the achievement levels in terms of percentages better than levels expressed as scale scores. When levels are expressed as percentages, people can evaluate whether a student's performance is near, at, above, or below the achievement level.

- NAEP achievement levels are used as a model in the development of district and State assessment standards.

- Achievement levels allow the tracking of different subgroups of students, especially with regard to gender and race.

- Achievement levels are used by policymakers, legislators, State boards of education, the press, and parents to suit their own purposes. For example, educators may use them to determine how the State is doing in comparison with another State or the entire Nation, and legislators may used them to advocate more funding for areas needing improvement.

- In some cases, achievement levels are not used to leverage funding decisions.

- Appropriate audiences for the achievement level results are all segments of the population.

Additional Comments

There was consensus within the group that parents and nontesting experts are more interested in looking at rankings than achievement levels. It was also noted that educators and teachers often ask for more detailed information, such as sample items, whereas less academic groups may be content with just the numbers that indicate student performance. Participants were asked whether they saw merit in moving away from reporting student performance in percentages and using a metric scale instead. For example, on a 100-item test the following total percentage of possible points might be reflective of students who scored at each level:

Basic	37%
Proficient	61%
Advanced	81.9%

Participants indicated that clarity about scores is important. It was noted that the percentage of total possible points could change each time the assessment test was given if the level of difficulty, the number of items, or the format of test items changed. Participants found the percentage score easier to understand without much explanation.

Because there are such large gaps between the scores, the group believed it was important for the public to know that student performance, as reflected in the achievement level scores, is determined using informed judgment rather than an arbitrary and capricious system.

Participants reiterated that it is difficult to determine the progress of students who score below Basic. There was further discussion on the regional clustering of scores and an acknowledgment that more homogeneous environments are likely to have more similar scores.

Attachment C

Discussion Group Participants: San Francisco, CA
November 22, 1999

San Francisco School Team
(Independent K–8 School)
300 Gaven Street
San Francisco, CA 94134
Phone: 415–239–5065

Ann Jaquith
Assistant Head of School
58 San Benito Way
San Francisco, CA 94127
Phone: 415–661–8061

Rebecca Greco
Teacher
1420 Seventh Avenue
San Francisco, CA 94122

Jane Nolan Yen
Parent
4745 17th Street
San Francisco, CA 94117
Phone: 415–661–8061

Junipera Sera Elementary School Team
School Team (K–5)
625 Holly Park Circle
San Francisco, CA 94110
Phone: 415–695–5685

Kevin Truitt
Principal
3642–A 19th Street
San Francisco, CA 94110
Phone: 415–552–4742

C.M. Whiteside
Teacher
448 Vicksburg Street
San Francisco, CA 94114
Phone: 415–282–5630

Ramon Martinez
Parent
446 30th Street
San Francisco, CA 94131
Phone: 415–641–9129

Francisco Middle School Team (6–8)
2190 Powell Street
San Francisco, CA 94133
Phone: 415–291–7900

Marian Seiki
Principal
2190 Powell Street
San Francisco, CA 94133
Phone: 415–291–7900

Gerald Pelletier
Assistant Principal
45 Hickory Road
Fairfax, CA 94930
Phone: 415–456–0974

Gentle Blythe
Parent Liaison
2190 Powell Street
San Francisco, CA 94133
Phone: 415–291–7900

Lowell High School Team (9–12)
1101 Eucalyptus Drive
San Francisco, CA 94132
Phone: 415–759–2730

John Mahoney
Assistant Principal
6 Driftwood Avenue
Novato, CA 94945
Phone: 415–898–1582

Steve Schmidt
Teacher
255 Buckingham Way, #802
San Francisco, CA 94132
Phone: 415–664–2957

Pam Olbrycht
Parent
50 Alviso Street
San Francisco, CA 94127

Report of Business Leaders

Greater Houston Partnership
Houston, Texas
December 9, 1999

Participant Sampling

NAGB held the last of its four discussion groups on December 9, 1999, in Houston, Texas. The staff of the Greater Houston Partnership recruited participants. The 10 participants included Texas business leaders, representatives of institutions of higher education who serve in an advisory capacity to the Greater Houston Partnership, and business leaders from the greater Houston area involved in school reform efforts. Attachment D to this report provides a list of participants.

The purpose of the Houston discussion group was to gather information on business leaders' perceptions on the reasonableness, usefulness, and utility of student achievement levels. This report summarizes the exchange of ideas that took place regarding these issues. To protect the confidentiality of group participants, the report does not attribute comments or ideas to specific individuals.

Results: Are Levels Reasonable?

The initial discussion concerning the reasonableness of achievement levels focused on three main points:

1. Policy definitions of achievement levels.
2. Content descriptions of achievement levels.
3. Relation of NAEP achievement levels to other assessments.

1. Policy Definitions of Achievement Levels

Participants were in general agreement that the use of achievement levels was justified, because they appeared capable of measuring increasingly difficult levels of material. For example, the definition of the Basic achievement level as representing partial mastery of subject matter was one that participants found entirely appropriate. Although their ubiquitous concern was about the numbers of students who score below the Basic level, group members agreed that the standards should not be lowered.

One participant noted, "If you have good standards, don't tinker with them." With that in mind, the group agreed that NAEP had set very high performance standards, but that it was up to the schools, their administrators, and their teachers to ensure that students are prepared to meet those standards. Another participant suggested keeping the standards where they are so that students can aspire to attain them. Student performance expectations should be raised rather than be lowered.

Participants believed there was no need to change the achievement levels of Basic, Proficient, and Advanced.

2. Content Descriptions of Achievement Levels

As a group, participants did not have any difficulties with the content descriptions. One participant suggested that the high expectations and disappointment over poor performances create a fervor over the validity of the tests. Participants noted that in many foreign countries, there is a high expectation of student performance, while in the United States there is an inclination to reduce the standard to a lower common denominator. Again, participants noted that expectations should be raised regarding student performance rather than reducing the expectation for performance on the assessments.

3. Relation of NAEP Achievement Levels to Other Assessments

Business leaders and education activists remarked that it is bad national policy to use the standardized score results from tests such as the SAT and ACT to compare students' performance on NAEP. This is because the tests have different purposes and students have different motivation factors for taking them.

Results: Are Levels Informative to the Public?

The second area of discussion focused on audience experience and reaction with regard to the achievement levels. Comments focused on two main areas:

1. Reporting of achievement levels.
2. Usefulness of achievement results.

1. Reporting of Achievement Levels

With regard to the process for reporting test results to the public, group members noted that percentages are easier for the public to understand than scale measurements. However, scale scores are good for showing improvement over time and, unlike percentages, facilitate cross-State comparisons.

2. Usefulness of Achievement Levels

Participants reported that achievement levels serve two purposes:

- Indicate the level of a student's achievement.

- Provide information about what knowledge a student should have absorbed and what level the student should be.

With regard to the overall usefulness of the achievement levels, there was consensus within the group that educators and parents use the assessments to measure students' success. Participants also made the following points about NAEP achievement levels:

- Achievement levels are easy to interpret and provide a common-sense approach to interpreting results.
- Achievement levels are assessable.
- The exemplars in the demonstration books are helpful in understanding the assessment.
- Trend data are useful, as is the consistency of States using the same data to determine their relative standing.
- Nonexperts can most easily understand the achievement levels when they are reported in terms of percentages, making it easier for people to evaluate whether a student's performance is near, at, above, or below the achievement level.
- In Texas, the NAEP achievement levels are used as an outside validation of Texas assessment standards.
- Achievement levels foster an awareness of the performance of different subgroups of students, especially with regard to gender and race.
- Achievement levels are often used by policymakers, legislators, State boards of education, the press, parents, and State education staff to suit their own purposes, such as when legislators use them to assess funding decisions on education.
- Achievement levels are not used to leverage funding decisions because, as mandated by law, all schools receive similar funding amounts.
- Achievement levels are used as a model for the State-developed criterion referenced assessment in Texas, but they cannot be used to evaluate the performance of a school district or to determine policy.

Additional Comments

Additional comments included the following:

- One participant questioned the decision to report the below Basic data in the NCES report.
- The results of 12th grade students may not truly reflect students' capabilities because there are no consequences related to students' scores. The NAEP assessment does not affect graduation or getting into college.
- There was discussion on the possibility of using a shortened version of the NAEP assessment as a State test in Texas, and how it could serve as a powerful tool for change.

- The group noted that NAEP is similar to the Voluntary National Tests (VNT) in that there is no requirement to participate in either assessment. Districts would be invited to participate and if they opted not to, individual schools within districts could request to participate. It was noted that urban districts have expressed interest in participating in the VNT, although lobbyists have indicated that urban districts may be the ones that would score most poorly on such assessments because of their diversity in terms of student needs and achievement.

References

National Academy of Sciences. (1998). *Grading the Nation's report card: Evaluating NAEP and transforming the assessment of educational progress.* Washington, DC.

Bourque, M.L., Champagne, A.B., and Crissman, S. (1997). *1996 science performance standards: Achievement results for the nation and the states*. Washington, DC: National Assessment Governing Board.

Attachment D

Discussion Group Participants: Houston, TX
December 9, 1999

Lucretia Ahrens
Manager, Community Relations
Reliant Energy
11 Louisiana
Houston, TX 77002

Beth Ann Bryan
Governor's Business Council
2901 Bammel Lane, Suite 29
Houston, TX 77098

John Cater
President
Compass Bank
24 Greenway Plaza, Suite 1400
Houston, TX 77046–2401

Linda Clarke
Executive Director
The Houston Annenberg Challenge
1001 Fannin, Suite 2210
Houston, TX 77002–6709

Ken DeDominicis
Vice President of Institutional Advancement
University of Saint Thomas
3800 Montrose Road
Houston, TX 77006–4626

Nancy Rose
Greater Houston Partnership
1200 Smith, Suite 700
Houston, TX 77002–4309

Marina Ballantyne Walne
Education Specialist
American Productivity & Quality Center
123 North Post Oak Lane, Third Floor
Houston, TX 77024–7797

Greg Weiher
Professor
Department of Political Science
University of Houston
4800 Calhoun
Houston, TX 77204–2162

Darv Winick
President
Winick & Associates
12402 Pine Oak Drive
Dickinson, TX 77539

NAGB Board Member
John Stevens
Executive Director
Texas Business and Education Coalition
400 West 15th Suite 910
Austin, TX 78701

NAGB Staff
Mary Lyn Bourque
Roy Truby

SECTION 9

Conclusions and Recommendations

Sheila Byrd Consultant

November 2000

Conclusions and Recommendations

Sheila Byrd

By commissioning the studies contained herein, the National Assessment Governing Board (NAGB) affirms its commitment to ongoing research and the role it plays in the development of the Board's evolving standards-setting and standards-reporting processes. At the beginning of its second decade, NAGB is considering not only the efficacy of its past policies and practices, as some of this research describes, but also the ways in which it may yet address the concerns of National Assessment of Educational Progress (NAEP) consumers without abrogating its commitment to high standards.

NAGB's policy statement presents the Board's understanding of the standards-setting process and the need for continuous evaluation of the process over time:

> The development of achievement levels requires vigilance to ensure that aspects of the level-setting process not be prematurely institutionalized, closing off new ways of thinking about the levels, new ways of expressing assessment frameworks in terms of the levels; and new technologies for assessing student performance, interpreting NAEP data, and reporting NAEP results.[1]

These studies attest to the fact that despite permutations in membership, NAGB has consistently maintained these principles, and they suggest that NAGB will continue to rely on them as it makes policy decisions precipitated by reports such as this.

The conclusions and recommendations that follow are organized in a way that allowed board members to address the recurring issues both within and across the various studies, but they are also designed to allow a broader audience to consider NAGB's standards-setting and standards-reporting processes. NAGB surely would like to secure continued public understanding of and support for NAEP, as well as of the role that NAGB plays in setting high standards for student achievement in America. Three simple questions provide a helpful framework for the improvements and policy issues to be considered:

1. What does NAGB do?

2. Does the general public understand NAEP results? Why or why not?

3. Are there policy changes NAGB should consider to help achieve its goals?

[1] National Assessment Governing Board (1993). *Developing Student Performance Levels for the National Assessment of Educational Progress.* (Policy Statement). Washington, DC: National Assessment Governing Board.

What Does NAGB Do?

Understanding why NAGB was established helps us understand what NAGB does, because its responsibility has not changed. Brown's[2] research reminds us that the National Assessment, before NAGB's creation, produced normative, national survey data for the professional community. In other words, data from the tests revealed only how well students performed relative to other students; it did not report how well they performed against a standard—desirable academic expectations appropriate for that age or grade level. "Members of NAGB believed that normative models could mislead the public," Brown observes, "if the average score of the national group was not reflective of sufficient quality" (p. 14). He later proposes for the board that the essential policy issue continues to be whether NAEP results should still be reported in terms of quality or return to being reported normatively (p. 38).

"NAGB was clearly troubled by reporting on the average score of the Nation as the referent of quality," Brown remarks. "NAGB believed that qualitative reporting would provide an impetus for change even if the performance levels were not satisfactory initially" (p. 38). The new process would evaluate student achievement based on the core knowledge of what students *should* know to be Proficient, and, although it was part of NAGB's legislated mandate, NAGB's decision to develop performance levels was still controversial.

Reporting by performance levels meant that NAEP would now codify for the public the proportions of students achieving at various levels: what percentage of students was at, below, and above Proficient. Brown also notes: "It was believed by NAGB that NAEP results reported on meaningful performance levels would be more understandable by the public and more useful to those who make instructional and policy decisions" (p. 14).

Brown's description of the various evaluations suggests that the levels were perhaps too understandable and calls to mind the truism that when folks don't like results, they scrutinize process. Having said that, however, it is clear that public scrutiny has served the standards-setting process well. By examining all of the external evaluations, many of which are criticisms of the model, process, and product of NAGB's original achievement levels setting, Brown is able to describe important points at which NAGB either affirmed or altered aspects of the process to make it better.

For example, critics have charged that the judgment tasks that panelists are asked to perform within the NAGB standards-setting model are conceptually too difficult. Reckase points out, however, that this criticism has been leveled only by those outside the process; past panelists report that the tasks are not too demanding.[3] Other critics have charged that the process was

[2] Brown, W. (2000). Reporting NAEP by Achievement Levels: An Analysis of Policy and External Reviews. In M. L. Bourque and S. Byrd (Eds.), *Student Performance Standards on the National Assessment of Educational Progress: Affirmation and Improvements* (pp. 13–39). Washington, DC: National Assessment Governing Board.
[3] Reckase, M.D. (2000). A Survey and Evaluation of Recently Developed Procedures for Setting Standards on Educational Tests. In M. L. Bourque and S. Byrd (Eds.), *Student Performance Standards on the National Assessment of Educational Progress: Affirmation and Improvements* (pp. 43–69). Washington, DC: National Assessment Governing Board.

poorly implemented in early years, that the training of panelists was inadequate or inconsistent, and that the panelists lacked technical expertise. Finally, some charge that the product provides an underestimate of students who are Proficient and Advanced and that it is not validated using traditional psychometric procedures.

Illuminating NAGB's response to each evaluation, Brown notes the extensive review and research conducted by ACT and the Technical Advisory Committee Meeting on Standard Setting. He describes the improved articulation among the frameworks, item development, and achievement level descriptions; improvements in the sampling plan; the extensive training of panelists; new forms of feedback; better matching of exemplar items and performance levels; and pilot studies.

The current NAGB process appears to be state of the art in the eyes of the researchers, as we also infer in the summary of State and commercial standards-setting efforts. Both State and commercial tests, according to Nellhaus and Forsyth, have come to reflect many aspects of the NAGB process, as have some of the alternative methods explored in the section by Reckase. It appears that none of the other methods (State, commercial, or proposed alternatives) have undergone the continuous evaluation and subsequent improvement that NAGB's has. Many alternatives described by Reckase have been used in limited research studies only, or merely described as possible procedures, thereby warranting further development before NAGB could apply them to NAEP. (See question 3 for further discussion of alternatives.)

Although the research indicates that NAGB may remain confident that its current process is the culmination of years of research and improvements, other studies herein indicate that some parents, educators, policymakers, and members of the public may be misunderstanding or misinterpreting the results.

Does the General Public Understand NAEP Results? Why or Why Not?

Despite the apparently successful evolution of the achievement level-setting process thus far, the research implies that not all parents, educators, policymakers, and the rest of the public understand the process nor the meaning of NAEP results as well as the Board may want. According to the Board's policy statement:

> The purpose for developing student performance levels on the NAEP is to clarify for all readers and users of NAEP data that these are expectations that stipulate what *students should know and be able to do* at each grade level and in each content area measured by NAEP. The achievement levels make the NAEP data more understandable to the general user, parents, policymakers, and educators alike. They are an effort to make NAEP part of the vocabulary of the general public.[4]

[4] National Assessment Governing Board policy.

Summaries of the focus group discussions, the analysis of the press coverage, and even the National Academy of Sciences' 1999 report[5] on NAEP acknowledge that the reporting of achievement levels has been popular and should continue. Still, there is evidence, particularly in the analysis of the press coverage and in Popham's study, that achievement levels and the reporting of results can be misunderstood or misinterpreted.

Brown's study notes how technical experts and some policymakers have responded over the years to NAGB's standards-setting process, in many cases sparking improvements. Hambleton's analysis of the press release and press clips focuses a slightly different lens on imminent policy considerations for NAGB with regard to reporting procedures. After evaluating the press packages from the past decade of NAEP releases, Hambleton concludes that they have improved considerably, and that newspapers do appear to be reporting the numbers from NAGB and the National Center for Education Statistics (NCES) correctly. He detects recurring problems, however, with how the media are reporting and interpreting NAEP results for the public, and he encourages NAGB to offer the press more and better training on the meaning of NAEP results.

The focus group discussions also seem to confirm that discrepancies in how the results are perceived still exist among various groups. In some cases it is an issue of State versus local coverage, and in others it may be the difference between the level of interest among education professionals and that of the public, whose background knowledge may be minimal. For example, policymakers and business leaders agreed that the policy definitions and content descriptions are reasonable, but the State assessment personnel and education professionals (not surprisingly, perhaps) tended to be more concerned, saying that the definitions and descriptions are laudable goals but are also value laden, subjective, and do not account for differences among local curriculums.

Hambleton's observations on "some problems in NAEP score reporting" may help explain such discrepancies:

> The achievement levels have generated some interest in NAEP scores but still do not appear to be fully understood. "Above Average" is substituted for "Advanced"; Basic students have been described as "basically competent." Language and examples need to be found to communicate the correct interpretations of Advanced, Proficient, Basic, and Below Basic. What are the knowledge and skills possessed by students at each level, and what are the differences among the performance categories? These appear to be two of the questions that need to be satisfactorily answered to improve the reporting of NAEP scores. (pp. 151–152)

[5] Pellegrino, James W., Lee R. Jones, and Karen J. Mitchell. (1999). Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress. Washington, DC: National Academy Press.

Focus group participants appear to agree that the use of exemplars, for example, is a good idea, although some of them believed the quality of the exemplars could be improved, as Hambleton's comment also implies.

Despite the fact that much of the statistical jargon has been deleted from the reports since 1994, reporters continue to use terms such as "statistical significance," albeit improperly, and they appear to have trouble understanding the scores and interpreting point differentials:

> Percentiles and cumulative percentages are two more statistical concepts with a history of being misinterpreted by the press. The confusion is passed on to the public. Also, the meaning of NAEP scores remains a problem. What is the meaning of a 1- to 3-point change, and how should a 1- to 3-point change be interpreted relative to a 5- to 8-point change? Unless ways can be found to interpret the scaled scores and scaled score differences, it may be safer not to report them. Benchmarking scaled scores could be helpful. For example, if the differences between boys and girls in a subject area is 5 points, then this becomes meaningful for judging the relative size of the differences among ethnic groups or changes in scores over time. (p. 152)

The meaning of scores is also an issue of concern to some policymakers, as Popham's analysis details. These users of NAEP data would probably say that not having enough information about student performance within levels is as problematic as misinterpreting the meaning of scores as currently reported. Changing the methodology for setting achievement levels and how scores are reported, however, are policy issues that are discussed in the next section, with regard to question 3.

The validity of the achievement levels themselves does not appear to be an issue in the press clips, and, for the moment, we are simply concerned about whether the public understands NAEP results as currently reported. There may be practical ways that NAGB and NCES can improve on the public's understanding of the standards without changing current policy. Hambleton's report implies the following such improvements that may be worthy of NAGB's consideration:

1. Continue to provide thorough, but lean data to the press. Too much complex information appears to cause reporters to make erroneous causal inferences, such as implying that watching too much television will cause lower NAEP scores (p. 138).
2. Explain more clearly the statistical concepts and scores. For example, provide more guidance about the meaning of the data (e.g., the gap in performance among racial groups) by tying the information about point differentials between groups explicitly to the differences among the groups, so that the public will make relevant connections and correct interpretations (pp. 137–138).
3. Continue the practice of employing exemplar items and good graphics in score reporting (p. 141).
4. Continue the use of "newspaperlike" reports, such as those for 1996 Science, 1998 Reading, and 1998 Writing (p. 149).

5. Select the most compelling findings, focus on interpreting those findings, and provide directions or remedies for improvement, where necessary (p. 150).

These recommendations build on communications work that NAGB and NCES have already accomplished.

In addition to continuing and improving these practices, however, the board has considered policy issues raised in the research. The next section addresses policy issues that emerge in the research of Popham and Reckase in particular.

Are There Policy Changes NAGB Should Consider To Help Achieve Its Goals?

This question goes beyond asking how NAGB can make its current standards-setting policies and reporting practices (and therefore the NAEP results) more understandable to the public. It asks whether NAGB should consider changing any of those policies. Does the research indicate that NAEP is *not* in fact achieving its desired goal of being an effective policy tool and an impetus for instructional reform? We have seen to the contrary that NAEP appears to have served as a model for the way in which both commercial test publishers and States have developed performance standards. We have also seen that press coverage has increased steadily since 1992. NAEP frameworks and test specifications are commonly used as benchmarks by those who develop content standards, curriculums, and assessments.

It is possible, however, that despite improvements in level-setting and reporting processes as currently formulated, misinterpretation of the results may still exist. NAGB's guiding principles suggested that the Board consider at least two policy issues that have emerged about the validity of the levels themselves:

1. The addition or reconfiguration of current achievement levels to allow more detailed data about performance within levels.
2. The piloting of alternative methodologies for setting the levels to explore their potential viability for use with NAEP.

At the November 1999 NAGB meeting, Secretary of Education Richard Riley suggested that the current achievement levels are "not as useful as I would hope they could be in terms of a person making public policy, whether it's a Governor, or a secretary, or superintendent, whatever,"[6] indicating that for at least some policymakers, NAEP is not achieving its stated goals.

(A) Addition or Reconfiguration of Current Achievement Levels

Riley suggested that NAGB consider adding an achievement level or somehow "convey where improvement is taking place or not taking place and where movement is happening." He stated that the Basic and Below Basic categories are very broad, whereas the Advanced category is "so very narrow that it's hardly useful."[7] Others may argue that, although it is disappointing that so

[6] Remarks of U.S. Secretary of Education Richard Riley, as quoted by Popham, p. 159.
[7] Ibid. (Popham, p. 160).

few students achieve the Advanced level, it does not necessarily follow that NAGB should begin to "curve" the results. The Popham study details various interpretations of NAGB achievement levels throughout the past decade and offers a succinct summary of the advantages and disadvantages for altering them.

In addition to discussing what some perceive to be the problematic nature of the achievement levels serving dual duty as both goals for, and measures of, student achievement, Popham highlights five modification options "that appear to be likely contenders for change." (pp. 174–175):

1. Add one or more achievement levels.
2. Divide the current levels into distinguishable, within-level reporting categories.
3. Make Below Basic a NAGB-sanctioned reporting category.
4. Relabel the existing achievement levels, especially Proficient.
5. Lower scale-score ranges associated with one or more achievement levels.

Option 1 (adding a level) "will do nothing to erase the continuing perception that hoards of American students are incapable of performing at a Proficient level," he notes. The second option, he argues, would not address the concerns of groups like the Council of Great City Schools, who fear that too many students will be classified in the Below Basic category. Option 4, relabeling the levels, although possibly changing negative characterizations of results, could be viewed as simply a disingenuous repackaging. Lowering scale scores, option 5, is still a "judgmental enterprise" that may be viewed by some as more realistic but by others as an obvious lowering of standards.

A hybrid of options 2 and 3 seems to emerge as the most viable way, according to Popham, for NAGB to respond to Secretary Riley's assertion that NAGB should "convey to the American people that yes, we have high standards, and none of us wants to toy with that—challenging standards is really an overall purpose of NAGB—but yes, we're also measuring improvement or the lack thereof in a useful way" (p. 160). Subdividing the Below Basic and Basic categories may require augmenting the number of items in the NAEP item pool "to make possible any meaningful within-level differentiation, particularly at the lowest and highest ends of the performance distributions" (p. 175), but doing so at the two "nongoal" categories[8] might satisfy those who want more information but not upset those who would not accept a lowering of standards.[9]

Finally, Popham suggests that six well-defined and exemplar-illustrated within-level subcategories (high, middle, and low) "should be carried out in the context of an *instructional* orientation that provides as much guidance as possible to the educators who are striving to help students reach higher levels of NAEP-assessed achievement (p. 178)." Although most members may agree that initiating a serious communications effort such as that described by Popham (pp. 180–181) may be a good idea, others believe that providing too much instructional guidance is beyond NAGB's scope of authority.

[8] As noted earlier, Popham advocates clarifying that *Proficient* is the goal.
[9] Such efforts to differentiate student performance have been tried at the state level (e.g., Kentucky).

The press analysis points out that 7 of 10 articles examined for the 1992 and 1996 Mathematics Assessments quoted Secretary Riley's suggestion that challenging curriculum, standards, and assessments can work to improve student performance. Popham suggests that NAGB consider focusing "more instructionally catalytic support for those educators who wish to employ NAEP's curricular frameworks and achievement subcategories in educative approaches that would benefit students (p. 180)."

The challenge for NAGB in this case would be to determine what "instructionally catalytic support" and support for standards, curriculums, and assessments would entail, and whether it is within NAGB's legislative mandate to offer it.

Achievement Levels Committee Deliberations

The Achievement Levels Committee, a standing committee of NAGB, chaired by Edward Haertel, met for 2 days in Snowbird, Utah, in June with the authors of the reports in this volume to discuss their recommendations and to prepare their own set of recommendations for the Board meeting in August 2000.

Although some members of the committee initially thought the notion of additional reporting categories as suggested by Popham might improve the communicability of the achievement levels, most did not agree. The Committee agreed, however, as did the Board, that the greatest need is to try to improve the way in which achievement levels are reported. Improving reporting might obviate some of the problems associated with communicability. They also agreed that it was imperative to try to reduce or eliminate misinterpretations of the levels. With respect to adding more reporting categories, the Board felt that the distinction between the *goals of what students should know and be able to do* and the reporting categories of what *students do know and are able to do* would too easily be lost. Consequently, the new reporting categories would become *de facto* new levels, thus leading to confusion rather than clarification in NAEP reporting.

They also agreed that research was needed on ways to improve the displays of what students know and can do on any given assessment (e.g., using item maps). Furthermore, the idea of developing reporting displays that show what students who are approaching Basic and approaching Proficient know and can do would greatly improve the understanding and communication value of the levels.[10] Finally, the Board agreed that examining alternative ways to report the performance of students in the Below Basic category might help reduce misinterpretation.

[10] See Figure 1 on p. 231 that displays a modified version of the Popham suggestion found on p. 179.

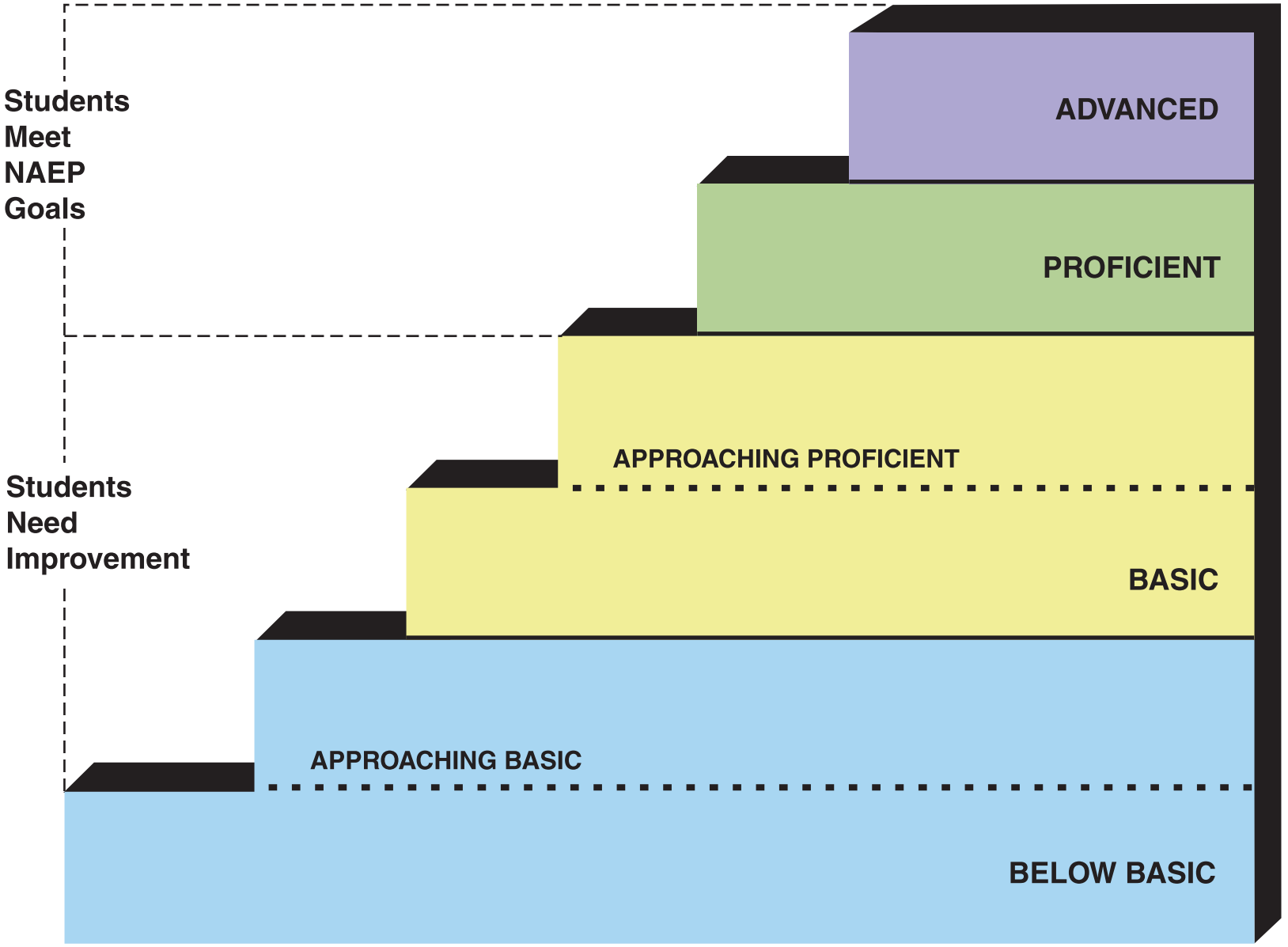


Figure 1. A recommended modification in NAEP reporting categories being researched by NAGB.

(B) Piloting of Alternative Methodologies for Setting Levels

The final policy question the board considered is actually twofold and is summarized nicely by Brown at the end of his paper:

1. Are the recommendations by critics to abandon the present achievement level-setting process warranted by the problems identified?
2. Is there a viable and tested model available that will produce results that are more valid and more reliable?

Brown's research, Popham's research, and board deliberations at its March 2000 meeting suggest that although past problems with the level-setting process probably do not warrant the abandonment of the process, there may be a way to provide more and better information about student achievement within the current levels without disturbing the stability of NAEP trends over time. "The policy decision to establish performance levels proved to be as technically complex as it was forward thinking," observes Dr. Brown. "During the past decade," he continues,

> NAGB has persisted with its policy decision, even in the face of considerable criticisms from noted psychometricians who labeled the achievement level-setting process as flawed. In response to criticism from reviewers and in search of improvements to the achievement level-setting process, NAGB continued to study alternate procedures that might improve the modified Angoff method that was in use . . . . The results of the research conducted by ACT have improved the standards-setting model considerably (p. 38).

Because NAGB is always committed to considering "new ways of thinking about the levels" and "new technologies for assessing student performance," as its policy statement avows, NAGB considered the recommendations offered by Reckase about the piloting of potential alternatives.

Reckase first acknowledges that any standards-setting effort is a judgmental process, and NAGB might do well to continue to emphasize that yes, the judgments are subjective, but they are the informed judgments of trained panelists. It is the Board that ultimately approves and takes responsibility for the standards. Its "highly evolved," systematic, and replicable process should perhaps obviate concerns, such as some voiced in the focus groups, that the "subjectivity" of the levels, in and of itself, is necessarily a bad thing.

Reckase has cataloged for the board the full range of alternative standards-setting methods and suggests the following criteria for evaluating the efficacy of alternative standards-setting methodologies:

- Judges can set the standard they intend.
- Tasks that judges are asked to perform are moderate in their cognitive complexity.
- Cut-scores have acceptable standard errors of estimate.
- Process is replicable.

One significant problem, however, is that many of the procedures suggested over the past decade have been used in limited research studies only or merely described as possible procedures. Reckase concludes that all would need extensive further development to connect the method to the policy and content frameworks, to develop methods for reporting results, and to withstand the types of public evaluations that have already been applied to NAGB's process, as we have seen.

NAGB Deliberations

After considerable discussion, the Board decided not to pilot an alternative method for setting standards. Their reasons were compelling. First, there is considerable agreement among the authors of this report that the current method is state of the art. There have been, over the past decade, vast amounts of research that have continued to improve the NAGB/ACT process for setting standards. The process used most recently in 1998 is considerably different from the method first used in 1992. The nearly decade of experience has refined the method to the point where there is ample evidence for the claim "state of the art." Second, if the Board were to pilot one or more methods proposed by Reckase as likely candidates, the Board would need to mount another decade of research to bring any new method up to the level of quality enjoyed by the NAGB/ACT process. Third, the adoption of a new method would signal a break in trends, because alternative methods generally would not yield comparable results to current methods. Finally, with limited resources, it was deemed more productive to work on issues of improving interpretation and minimizing misinterpretations.

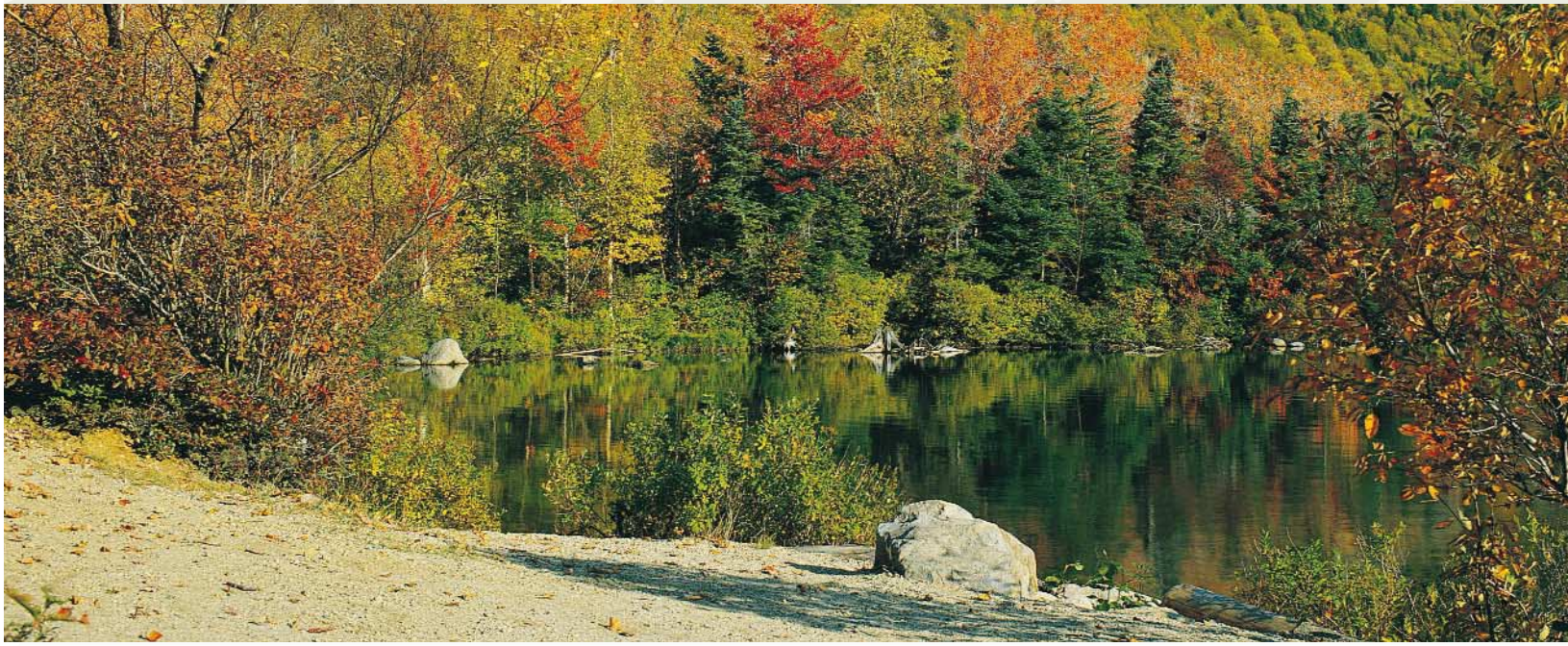
The Board will continue to direct its efforts to improve the current method. In conjunction with that goal, the Board advocates a continuing program of research on standard setting with a priority on validation, refinement of the methodology, and improving methods of reporting.

To summarize, the research indicates that NAGB can track significant progress over the past decade toward its goal of providing reasonable, valid, and informative expectations for what students should know and be able to do. There is evidence that the achievement levels have made the NAEP data more understandable to the public and that NAGB should continue to employ its improving communications strategies. Because there is still some concern over the validity of the levels, NAGB should consider modifying the levels to provide more information about students' progress toward reaching proficiency and continue to refine the current methodology for its potential ability to improve an already state-of-the-art process.

SECTION 10

Acknowledgments and Appendixes

November 2000



Acknowledgments

The Board thanks the authors whose work appears in this volume. Their informative presentations and learned papers provide a collection of some of the best and current thinking about achievement levels. Some of the authors have provided advice to the National Assessment Governing Board (NAGB) for many years. We are particularly grateful for their continued dedication, commitment, and professional support for this all-important work of the Board.

We would especially like to thank Aspen Systems, Rockville, MD; Sheila Byrd, Consultant, Bethesda, MD; William Brown, Brownstar, Inc., Cary, North Carolina; Robert A. Forsyth, University of Iowa; Ronald K. Hambleton, University of Massachusetts at Amherst; Jeffrey M. Nellhaus, Massachusetts Department of Education; W. James Popham, University of California at Los Angeles; and Mark D. Reckase, Michigan State University.

We also thank the participants in the focus groups conducted by Aspen in various locations across the country. In addition, we are grateful to the many attendees at the various meetings of the Achievement Levels Committee who offered their ideas and suggestions for ways to improve the NAEP standards-based reporting system using student performance levels. Their valuable comments and insights will undoubtedly influence the future of the NAEP standards.

Finally, special thanks go to Munira Mwalimu and the staff at Aspen Systems for their logistical and publications support and to Tessa Campbell of the NAGB staff for her program support role to the committee and the NAGB staff and for assisting the editors in preparing this report.

Appendix A
Committee Agenda

National Assessment Governing Board
Achievement Levels Committee
June 23–24, 2000

Cliff Lodge
Snowbird, Utah
801–933–2233 Voice 801–933–2264 FAX

Friday, June 23

Noon	Opening, welcome remarks (Lunch) Edward Haertel, Chair	Maybird Floor B
1 p.m.	***Topic 1: Achievement Level Descriptions (ALDs)*** Discussants: Mark Reckase Robert Forsyth	
	What purpose(s) do *preliminary* ALDs serve? What is their relation to framework? To content domain? Are they necessary? Should the format be standardized? What is their role in developing the standards? What purpose(s) do *final* ALDs serve? Do these purposes conflict? What is their relationship to exemplars and cut scores?	
2 p.m.	***Topic 2: NAEP in relation to state and commercial tests*** Discussants: Robert Forsyth Jeff Nellhaus	
	What can be learned from the states? What can be learned from the test publishers?	
3 p.m.	Break	

3:15 p.m.	***Topic 3: Reporting ALs with Exemplars, Cut Scores, ALDs*** Discussants: Ronald Hambleton James Popham How do ALs improve the "form and use" of NAEP? What appropriate inferences can be made? What changes should be considered?	
4:30 p.m.	Adjourn	

Saturday, June 24

7–8 a.m.	Continental Breakfast	Maybird Floor B
8 a.m.	***Topic 4: Policy Issues*** Discussants: William Brown James Popham What criticisms must be addressed in next decade? Are the policy definitions appropriate for the next decade? Is broad representation among panelists a desirable goal? How will we know when to "update as necessary?"	
9 a.m.	***Topic 5: Methodology*** Discussants: Mark Reckase Ronald Hambleton What new methodologies have potential for NAEP? What are the advantages/disadvantages?	
10:15 a.m.	Break	
10:30 a.m.	***Next Steps and Recommendations to the Board*** Edward Haertel, Chair	
Noon	Break	
12:15 p.m.	Lunch	
1 p.m.	Invited Public Forum	
4:30 p.m.	Adjourn	

Appendix B
List of Participants

National Assessment Governing Board
Achievement Levels Committee Meeting
Friday, June 23, 2000

Board Members

Edward Haertel, Chair
Mark Musick
Deborah Voltz
Michael Nettles

NAGB Staff

Mary Lyn Bourque
Larry Feinberg
Sharif Shakrani
Roy Truby

Consultants

William Brown
Brownstar, Inc.
Cary, North Carolina

Sheila Byrd
Sheila Byrd, Inc.

Robert Forsyth
University of Iowa

Ronald Hambleton
University of Massachusetts at Amherst

Jeffrey Nellhaus
Massachusetts Department of Education

W. James Popham
University of California at Los Angeles

Mark Reckase
Michigan State University

Others

Susan Loomis
NAEP Project Director
ACT, Inc.

Andrew Kolstad
National Center for Education Statistics

National Assessment Governing Board
Achievement Levels Public Forum
Saturday, June 24, 2000

Linda H. Frazer
Director, Validation and Research
Kentucky Department of Education
500 Meo Street
18th Floor, Capitol Plaza Tower
Frankfort, KY 40601

Marilyn Howard
Superintendent of Public Instruction
Idaho Department of Education
Box 83720
Boise, ID 83720–0027

Pasquale DeVito
Director, Office of Assessment
RI Department of Education
255 Westminster Street
Providence, RI 02903–3400

Edward Slawski
Harcourt Brace Educational Measurement
555 Academic Court
San Antonio, TX 78204–2498

Carol White
Education Associate
Delaware Department of Education
P.O. Box 1402
Dover, DE 19903–1402

Wayne Martin
Council of Chief State School Officers
One Massachusetts Ave.
Suite 700
Washington, DC 20001

Louis M. Fabrizio
Director, Accountability Services
North Carolina Department
of Public Instruction
301 N. Wilmington Street
Raleigh, NC 27601–2825

Judith Keonig
National Research Council
2001 Wisconsin Ave. HA 178
Washington, DC 20007


NATIONAL ASSESSMENT
THE NATION'S
REPORT
CARD
GOVERNING BOARD