DOCUMENT RESUME

ED 359 268 TM 020 113

TITLE	Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations. Report to Congressional Requesters.
INSTITUTION	General Accounting Office, Washington, DC. Program Evaluation and Methodology Div.
REPORT NO	GAO/PEMD-93-12
PUB DATE	Jun 93
NOTE	120p.
AVAILABLE FROM	U.S. General Accounting Office, P.O. Box 6015, Gaithersburg, MD 20884-6015 (first copy free; $2 for additional copies; orders for 100 or more to one address are discounted 25 percent).
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Academic Achievement; *Academic Standards; Educational Policy; Elementary Secondary Education; Evaluation Methods; Evaluators; *Mathematics Achievement; *Measurement Techniques; National Surveys; Research Problems; *Scores; Student Evaluation; Test Interpretation; Test Results; *Test Validity
IDENTIFIERS	General Accounting Office; *National Assessment Governing Board; National Assessment of Educational Progress; *Standard Setting

ABSTRACT

In September 1991, the National Assessment Governing Board (NAGB) announced standards for basic, proficient, and advanced achievement in mathematics and reported that few American students had reached these standards. Expert reviewers noted technical problems with the NAGB approach and questioned its results. In this report, the NAGB standard-setting approach and ability to provide policy guidance to the National Assessment of Education Progress (NAEP) are examined. The NAEP test-score standards set in 1990 were evaluated by examining the adequacy of item-judgment procedures and by studying whether the evidence supported NAGB's interpretation of the NAEP scores selected for each level. The investigation found that the standard-setting approach was procedurally flawed, and that the interpretations of the resulting NAEP scores were of doubtful validity. The NAGB improved its procedures substantially in 1992, but the issue of the validity of interpretation remains. The report concludes that the NAGB approach is unsuited for the NAEP. Alternative approaches are reviewed, but it is apparent that their use will be difficult as the NAEP is currently designed. Specific recommendations are given to help implement these alternative approaches. Six tables and three figures illustrate the discussion. Appendixes include comments from the U.S. Department of Education and from the NAGB, a summary description of the NAEP and other supplementary materials. A four-part bibliography is provided. Contains 44 references. (SLD)

United States General Accounting Office

GAO

Report to Congressional Requesters

June 1993

EDUCATIONAL ACHIEVEMENT STANDARDS

NAGB's Approach Yields Misleading Interpretations

GAO/PEMD-93-12

GAO

United States
General Accounting Office
Washington, D.C. 20548

Program Evaluation and
Methodology Division

B-251957

June 23, 1993

The Honorable William D. Ford
Chairman, Committee on Education and Labor
House of Representatives

The Honorable Dale E. Kildee
Chairman, Subcommittee on Elementary,
Secondary, and Vocational Education
Committee on Education and Labor
House of Representatives

In your letter of October 7, 1991, you asked us to review the approach by which the National Assessment Governing Board had established standards for student performance on the National Assessment of Educational Progress in mathematics and to report whether evaluators' concerns about the approach were warranted. We issued an interim response to your letter in correspondence dated March 11, 1992 (GAO/PEMD-92-22R). This final report includes a more extensive analysis of the validity issues raised by NAGB's approach, an examination of alternative approaches to setting standards for performance on the assessment, and a review of NAGB's capacity to provide sound guidance on technical issues. We recommend that the question of how to set and interpret performance standards be reopened and that structures and procedures for governing the assessment be reviewed to ensure that policies are technically sound as well as responsive to constituent interests.

We will send copies of the report to the Secretary of Education, the Commissioner of Education Statistics, and the Chairman and Executive Director of the National Assessment Governing Board. We will also send copies to others who are interested on request.

If you have any questions or would like additional information, please call me at (202) 512-2900 or Robert L. York, Director of Program Evaluation in Human Services Areas, at (202) 512-5885. Other major contributors are listed in appendix VII.

Eleanor Chelimsky
Assistant Comptroller General

Executive Summary

Purpose

In September 1991, the National Assessment Governing Board (NAGB) announced standards for basic, proficient, and advanced achievement in mathematics as measured by the National Assessment of Educational Progress (NAEP) and reported that few American students had reached these standards. This finding resulted from an approach to standard-setting that had several novel features. Expert reviewers noted technical problems with the approach and questioned its results. NAGB acknowledged that its procedures were imperfect but considered the results sufficiently sound to publish and the approach sufficiently promising to be mandated as the primary basis of all future NAEP reporting.

The question of how to set standards for educational achievement and measure progress toward them is currently of great interest, and NAGB's approach may serve as a model for other efforts. In view of the controversy surrounding this approach, the chairmen of the House Education and Labor Committee and the Subcommittee on Elementary, Secondary, and Vocational Education asked GAO to evaluate (1) its strengths and weaknesses, (2) its suitability and that of alternative approaches for use with NAEP, and (3) NAGB's capability to provide technically sound policy guidance to NAEP.

Background

Funded by the Department of Education, administered by the National Center for Education Statistics (NCES), and implemented by a technical contractor, NAEP tests American students in basic subjects every few years and estimates student achievement at the national level based on complex statistical techniques. NAEP's statutory purposes are to describe achievement and to track changes over time. For the past two decades, NAEP's results have been reported without reference to any goals or standards of how students ought to perform.

In 1988, the Congress created NAGB, an independent and broadly representative governing board, to provide policy guidance for the assessment. The 1988 law also made NAGB responsible for identifying appropriate achievement goals for each subject and grade tested. In the hope of interpreting NAEP results in terms of standards for what students should know and be able to do, NAGB mandated a standard-setting approach that included (1) defining three levels of achievement in general terms, (2) using expert panelists to judge how students at each level should do on each item on the NAEP mathematics test, (3) selecting a NAEP score to represent the lower border of each level, and (4) interpreting performance at these scores in terms of the definitions and of statements

of what students at each level should be able to do. NAGB applied this approach to the 1990 NAEP mathematics test on a trial basis and to mathematics, reading, and writing in 1992.

GAO evaluated the NAEP test-score standards NAGB set in 1990 by examining the adequacy of NAGB's item judgment procedures and whether evidence supported NAGB's interpretation of the NAEP scores selected for each level. GAO also identified alternative standard-setting approaches and analyzed them to find which would work with the NAEP test as it is now designed. Lastly, GAO reviewed how NAGB made key decisions, especially how it used technical advice and information, in the level-setting case and two others.

Results in Brief

GAO found that NAGB's 1990 standard-setting approach was procedurally flawed and that the interpretations that NAGB gave to the resulting NAEP scores were of doubtful validity. While the scores selected represent moderate, strong, and outstanding performance on the test as a whole, GAO concluded that they do not necessarily imply that students have achieved the item mastery or readiness for future life, work, and study specified in NAGB's definitions and descriptions. The difficulties evident in NAGB's 1990 achievement levels resulted in part from procedural problems but also from the effort to set standards of overall performance (how good is good enough) that would also represent standards of mastery (what students at each level should know and be able to do). NAGB improved its standard-setting procedures substantially in 1992, but the critical issue of validity of interpretation—an issue in NAGB's approach—remains unresolved. GAO therefore concluded that NAGB's approach is unsuited for NAEP.

GAO identified several alternative approaches that could be used to establish standards for overall performance on a NAEP test. However, any approach that sets standards purporting to measure mastery of particular subject content will be difficult to use with NAEP as it is currently designed.

GAO found that in the case of the achievement levels, NAGB designed and implemented its approach without adequate technical information. In two other cases, however, NAGB made better use of such information. GAO concluded that NAGB's composition, procedures, and relationships with the Department of Education are inadequate to ensure that policy guidance to NAEP will be technically sound.

Principal Findings

Problems in NAGB's Approach

NAGB based its approach on a well-known standard-setting method but modified this method in untested ways. GAO found NAGB's 1990 procedure unusual in three respects: (1) the achievement levels are intended to reflect mastery of different types of materials, not merely differences in overall performance; (2) panelists applied their own individual ideas of the mathematics skills pertinent to each level rather than using a consensus-based standard; and (3) the panelists were not assisted in making informed judgments of how students who met the expectations for lower levels actually would perform on very difficult items. These departures left panelists' individual views—not informed consensus—as the basis for NAGB's standards.

Although there was reason to think that it might be difficult to translate item judgment results into a NAEP score at which students show both the overall performance and the type of mastery expected for each achievement level, NAGB did not compare actual to expected performance at the scores it selected, nor did it seek evidence that the interpretations it gave to those scores were valid compared to other evidence of student performance. Finally, NAGB presented its findings without advising readers that their validity had not been established and that problems of reliability had been found.

Alternative Standard-Setting Methods

GAO found that the NAEP scale can be used to express standards for overall performance on grade-level materials. There are several methods for setting such standards; each combines judgments about desirable levels of performance with data about what students at various NAEP scores are able to do. Because of the way NAEP is now designed, the current NAEP scale score is not a good way to measure students' knowledge of specific areas of school content, especially not advanced material that few students are taught. If such measurement is desired, new tests will probably need to be developed, or NAEP will need to be redesigned.

NAGB's Capabilities

GAO concluded that NAGB's strength lies in its broad representation, not in its technical expertise. However, the law assigns NAGB responsibility for some functions that are clearly technical and for others that have both technical and policy implications. From examining three decisions, GAO

found that when NAGB recognized an issue as clearly technical, it sought and used expert technical advice in policy planning and sometimes in implementation. However, NAGB initially considered the setting of achievement levels a policy function that it itself could perform with minimal technical support and did not appreciate the importance of verifying the validity of its score interpretations. NAGB's governance structure and procedures neither ensure that technical issues will be recognized nor require that technical considerations be addressed early in the policy formation process. GAO thus concluded that there is substantial continuing risk that NAGB may give NAEP technically unsound policy direction.

Recommendations

Since the current NAGB approach to setting standards has yielded unsupported interpretations of NAEP scores, GAO recommends (1) that NAGB withdraw its instructions to NCES to publish 1992 NAEP results primarily in terms of levels of achievement, (2) that NAGB and NCES review the achievement levels approach, and (3) that they examine alternative approaches.

To strengthen NAGB's capacity to give sound policy direction, GAO recommends that NAGB (1) obtain NCES review of proposed policies; (2) conform to its own policy of prescribing policy ends, not technical details; and (3) nominate for the testing and measurement positions on NAGB persons who are trained in the design and analysis of large-scale educational tests. GAO also recommends that the Congress clarify what it intends NAGB to do with respect to achievement goals and review the division of responsibilities between NAGB and NCES, with a view toward concentrating NAGB's efforts on the representational functions for which it is well designed.

Agency Comments

GAO requested and received comments from both NAGB and the Department of Education (for NCES). The department generally concurred with GAO's findings and recommendations. NAGB generally took issue with GAO's analysis and conclusions and with some of the recommendations as well. The full text of these comments and GAO's responses to them are in appendixes I and II.

Contents

Bibliography

Related GAO Products

Tables

Figures

Abbreviations

ACT	American College Testing
ETS	Educational Testing Service
GAO	General Accounting Office
GED	General Educational Development
IAEP	International Assessment of Educational Progress
NAE	National Academy of Education
NAEP	National Assessment of Educational Progress
NAGB	National Assessment Governing Board
NCES	National Center for Education Statistics
NCEST	National Council on Educational Standards and Testing
SCANS	Secretary's Commission on Achieving Necessary Skills
TSA	Trial State Assessment

Introduction

In 1990, the National Assessment Governing Board (NAGB) undertook a pioneering effort to set standards for student performance on the federally sponsored National Assessment of Educational Progress (NAEP). NAEP periodically tests national samples of students in grades 4, 8, and 12, describes their achievement in basic academic subjects, and analyzes how achievement has changed over time. Legislation enacted in 1988 directed NAGB to identify achievement goals for each grade and subject tested. NAGB responded by designing a standard-setting approach that defined three levels of achievement (basic, proficient, and advanced), identified a test score to serve as a performance standard for each level, and described what students who meet each standard should know and be able to do.

NAGB applied its approach to the 1990 NAEP assessment in mathematics, a subject area that has recently been of great concern and one that received special attention in the national education goals adopted by the president and the nation's governors in 1989. Using its performance standards to analyze 1990 NAEP test results, NAGB found that over one third of the students in each grade did not reach even the basic level of achievement in mathematics, which connotes partial mastery of fundamental skills. According to NAGB's analysis, just under half of the students tested in each grade had reached the basic level but could not be considered proficient—had not mastered challenging material—for their grade. Between 15 and 19 percent had scored at the proficient level or higher, and only 1 to 3 percent scored high enough to be considered advanced.[1]

These findings received wide attention. The National Education Goals Panel, a bipartisan association of governors, senior national administration officials, and congressional representatives established to monitor and report progress toward the national educational goals, incorporated NAGB's findings into its first report. Using the performance standard for the proficient level as its benchmark, the panel interpreted NAGB's findings to mean that less than one student in five had attained the national goal of demonstrating competency in mathematics.[2]

Both NAGB's approach and the findings that flowed from it have been controversial. Some users of NAEP data have applauded NAGB for providing standards where there had been none. Experts in testing and measurement, however, have noted possible flaws in NAGB's approach that

[1]National Assessment Governing Board, The Levels of Mathematics Achievement, vol. 1, National and State Summaries (Washington, D.C.: 1991), p. 34.

[2]National Education Goals Panel, The National Education Goals Report: Building a Nation of Learners, 1991 (Washington, D.C.: 1991), p. 12.

could lead to the selection of NAEP scores that do not accurately represent the three achievement levels and have recommended that the approach be reexamined.

The chairmen of the House Committee on Education and Labor and the Subcommittee on Elementary, Secondary, and Vocational Education asked us to conduct an independent evaluation of NAGB's approach to setting standards through NAEP. As the request letter noted, the question of how to set standards for educational achievement and measure progress toward them is currently of great interest, and NAGB's approach may serve as a model for other efforts. It is therefore important that both its strengths and possible weaknesses be identified and made public. Moreover, NAEP will undoubtedly play a central role in a system of assessments aligned to national content standards such as was recommended by the congressionally mandated National Council on Education Standards and Testing (NCEST). As the NCEST report observes, it is critical to ensure that assessments support valid, reliable, and fair measurement of the standards and that students are protected against unintended negative consequences of assessment approaches that are still being refined.[3]

Background

NAEP has measured student achievement in core academic subjects every few years since 1969. Initially funded by grants, the assessment was authorized by statute in 1978 and was added to the responsibilities of the National Center for Education Statistics (NCES), within what is now the Department of Education. NCES carries out the assessment with the assistance of a technical contractor, currently the Educational Testing Service (ETS). Until 1990, NAEP tested a nationally representative sample of students and reported only national and regional results. NAEP assessments are designed by broad-based consensus groups and emphasize material that is commonly taught for each grade and subject tested. (Summary information about NAEP can be found in appendix III.) Traditionally, NAEP reports have simply described what students can do; they have not prescribed norms or standards of what students should be able to do.

In 1988, amendments to NAEP's authorizing statute expanded the assessment to include state-level testing and established a new governance structure (shown in appendix IV). Under this new structure, the

[3]National Council on Education Standards and Testing, Raising Standards for American Education (Washington, D.C.: 1992). NAEP does not report results for, and thus cannot have consequences for, individual students. However, NAEP data may contribute to decisions about educational policy. If NAEP cannot adequately measure performance against educational standards or if it does so inaccurately, these decisions could be misguided.

Commissioner of Education Statistics, who heads NCES, retains responsibility for NAEP operations and for technical quality control. NAGB, a governing board appointed by the Secretary of Education but independent of the department, provides policy guidance for NAEP. NAGB's composition is set by statute. Its 23 members include governors and other state officials, district officials, teachers, principals, noneducators, and two testing and measurement experts.

In addition to providing policy guidance, NAGB is responsible for specific functions formerly performed by a panel that advised the NAEP contractor, including selecting the subject areas to be addressed, ensuring that each assessment's content is planned through a national consensus, and developing guidelines for reporting. The 1988 amendments also gave NAGB a special responsibility, that of "identifying appropriate achievement goals" for each subject and grade tested.

NAGB's Approach to Identifying Achievement Goals

The legislative record provided no guidance to assist NAGB in interpreting congressional intent with respect to the responsibility to identify achievement goals. Educational practice offered little guidance, either: the idea of setting goals or standards for student performance on a broad-based assessment like NAEP (as opposed to passing scores on tests of individual achievement) was relatively new. There was no established meaning for the term "achievement goal." Indeed, achievement goals or standards in education might be interpreted to mean

- content standards that identify what students should know,
- performance standards that identify the levels of performance that students should achieve, and
- performance targets that identify the percentage of students who should meet a performance standard.

In developing its approach, then, NAGB had to decide what kind of goal it wanted to establish and then consider how this might best be done.

Development of NAGB's Approach

NAGB's initial discussions, early in 1989, focused on how future NAEP tests might be designed to reflect content standards as well as current educational practice. As events proceeded, NAGB's attention shifted to the question of how content or performance standards might be established by means of tests that had already been designed or administered.[4]

[4]Tests to be administered in 1990 had already been designed. The reading and writing tests were to be redesigned for 1992, but the mathematics test design (new in 1990) was scheduled to be used through 1994.

Standard-setting based on existing tests, if feasible, would enable NAEP to measure progress toward the national education goals as early as 1990.

NAGB conducted a preliminary review of standard-setting methods and proposed in December 1989 that item judgment procedures might be used to set a single performance standard that represented adequate mastery of "core" content for each grade tested in the 1990 assessment of mathematics.[5] The proposed method would identify what students should know—that is, it would set a content standard. In addition, it would locate the test score that represents qualified performance with respect to that content—a performance standard. NAGB sought public and expert comment on the concept paper outlining this plan.

After reviewing the comments, NAGB concluded that an item judgment procedure known as the Angoff method could be applied to NAEP but that three performance levels or standards per grade would be needed: a challenging standard of proficiency; a lower, basic standard to direct attention toward students with the greatest need for improvement; and a standard for "world class" advanced performance. In May 1990, NAGB adopted a policy that defined three levels of achievement and specified that modified Angoff item judgment procedures be used to set three performance standards for each level and grade—to find the threshold score on the 500-point NAEP scale at which the criteria for each achievement level were met—beginning with the 1990 mathematics assessment on a trial basis.

Key Steps in the Approach

The definitions of basic, proficient, and advanced achievement, which reflect the language of the national educational goals and are intended to be applicable to every subject and grade tested by NAEP, form the foundation of NAGB's approach. NAGB's definitions are

"Basic. This level, below proficient, denotes partial mastery of knowledge and skills that are fundamental for proficient work at each grade level—4, 8, and 12. For 12th grade, this is higher than minimum competency skills (which normally are taught in elementary and junior high schools) and covers significant elements of standard high-school-level work.

"Proficient. This central level represents solid academic performance for each grade tested—4, 8, and 12. It reflects a consensus that students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. At grade 12, the proficient level encompasses a body of

[5]In an item judgment procedure, judges estimate how students who have the capabilities needed to meet a given standard would perform on each item on the test. The judgments are cumulated across items to form a total score. NAGB's item judgment procedure is described below.

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.

"Advanced. This higher level signifies superior performance beyond proficient grade-level mastery at grades 4, 8, and 12. For 12th grade, the advanced level shows 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."[6]

Table 1.1 shows the steps through which NAGB translated these definitions into NAEP scores and interpretations of those scores, with illustrations drawn from NAGB's procedures with respect to the 1990 8th grade basic level.[7] In step 2, item judgments, NAGB convened panels of teachers and other experts and asked them to judge how students who just reached the basic, proficient, or advanced level should perform on each item on the 4th grade, 8th grade, or 12th grade test. Next (step 3), NAGB combined panelists' judgments to calculate the percentage of items on each test that should be answered correctly by students at the lower margin of each achievement level.[8]

[6]National Assessment Governing Board, p. 5.

[7]Our purpose here is simply to describe. We examine the logic of NAGB's approach and the validity of its results in chapter 2.

[8]To illustrate, suppose that there are five questions on the test. Panelist A expects that 55, 80, 65, 20, and 70 percent of marginally basic students should give a correct answer to questions 1, 2, 3, 4, and 5, respectively. The average of these five estimates is 58. Fifty-eight percent correct is panelist A's estimate of the percent correct performance standard that a marginally basic student should achieve on this five-item test. Suppose that panelist B estimated a 46-percent standard and panelist C estimated 53 percent. The average for all three panelists, 52, would be the percent correct performance standard for basic achievement.

Table 1.1: The NAGB Approach: An Illustration

Step	8th grade basic level
1. Level definition	"This level, below proficient, denotes partial mastery of knowledge and skills that are fundamental for proficient work at each grade level"
2. Item judgments	Drawing on the level definition, panelists examined each item on the 8th grade test and judged how many marginally basic students out of 100 should be expected to answer that item correctly
3. Percent correct performance standard	Each panelist's item judgments were averaged; each panelist adjusted his or her average to represent a "best guess" of what the standard should be; the average best guess across panelists was computed, resulting in a performance standard of 48 percent correct
4. NAEP score performance standard	The NAEP score of 255, which corresponds to 48-percent correct, was selected as the lower threshold of the basic level
5. Percentage of students who met the NAEP standard	1990 NAEP results showed scores of 255 or above for 62.1 percent of 8th graders
6. Illustrative items: expected performance	Items that judges expected 80 percent of marginally basic students to answer correctly were classed as "basic" items
7. Illustrative items: actual 1990 performance	Items actually answered correctly by 80 percent of the students who scored at or near the basic standard of 255 (that is, between 242.5 and 267.5) on the 1990 assessment were classed as "basic" items
8. Achievement level description (statement of expected mastery) summarizing items common to steps 6 and 7	"BASIC: Partial Mastery of Knowledge and Skills. The eighth-grade student performing at the basic level should be able to identify and use the correct operations for solving one- and two-step problems involving addition, subtraction, multiplication, and division of whole numbers and decimals. . . "

Source: National Assessment Governing Board, The Levels of Mathematics Achievement, vol. 3, Technical Report (Washington, D.C.: 1991), pp. 13-14, 32-33, 58-61, 68, and 334.

NAGB then asked the NAEP technical contractor, ETS, to translate the percent correct standard into an equivalent score on the 500-point NAEP proficiency scale for the 1990 mathematics test (step 4) and to calculate the proportion of U.S. students who met or exceeded each such score based on the 1990 test data (step 5).[9] NAGB's results to this point are shown in table 1.2. The percent correct standards for the 4th, 8th, and 12th grades were very similar to one another, as were the percentages of students who scored at or above the basic, proficient, and advanced level. Just over 60 percent of the students in each grade were found to have reached the

[9]The NAEP scale covers the full range of proficiency tested, from the least-proficient score possible for 4th graders to the most-proficient score possible for 12th graders.

basic level. (The 60 percent includes students who reached the two higher levels.) Fifteen to 18 percent reached at least the proficient level, of which a handful achieved sufficiently high scores to be classified as advanced. Scores for nearly 40 percent of the students in each grade fell below the standard for the basic level.[10]

Table 1.2: Summary of NAGB Results for 1990[a]

Grade and achievement level	Percent correct performance standard	NAEP score performance standard	Percentage of students who scored at or above the NAEP standard
4th grade			
Basic	45%	207	63.3%
Proficient	68	245	14.9
Advanced	87	283	0.6
8th grade			
Basic	48	255	62.1
Proficient	72	295	18.1
Advanced	89	336	1.0
12th grade			
Basic	47	282	64.4
Proficient	73	330	16.2
Advanced	88	358	2.6

[a]The achievement level results for 1990 (shown here) were revised in 1992. The revised figures apply the standards adopted through the 1992 standard-setting process to the 1990 test, taking into account differences in test composition for the two years. The revised figures are included in U.S. Department of Education, Office of Educational Research and Improvement, NAEP 1992 Mathematics Report Card for the Nation and the States (Washington, D.C.: April 1993).

Source: National Assessment Governing Board, The Levels of Mathematics Achievement, vol. 1, National and State Summaries (Washington, D.C.: 1991).

To help users of NAEP data interpret performance at each score standard, NAGB identified items that the item judgment panelists expected most students at each level to answer correctly (step 6) and items that most students who scored at each NAEP performance standard actually did answer correctly (step 7). Next (step 8), NAGB created summary paragraphs based on items that students should and did answer. Termed "achievement levels descriptions," these paragraphs illustrate what students at each level should know and be able to do on the NAEP test. NAGB's 1990 achievement level descriptions for mathematics are printed in appendix V.

[10]NAGB labeled these scores "below basic" but did not define or interpret "below basic" achievement.

NAGB's approach is based on an item judgment method first proposed by William Angoff, a specialist in testing at ETS. The Angoff method has been widely used to set standards for individual performance on tests and is generally thought to be the most practical of the item judgment methods. Typically, the procedure is used to set a single standard or passing score that can be used to distinguish individuals who are qualified for some purpose (such as course credit, licensure, certification, or entry into a program of study) from those who are not qualified, on a test designed for that purpose. NAGB's approach modified standard Angoff procedure to set standards of how students at three achievement levels should perform—not on a test that was designed to measure individual performance at these levels but, rather, on one designed to describe proficiency for the overall student population. Thus, it employed a well-established approach but applied it for a new purpose and in a new way.

Implementation of the NAGB Approach

Early in the summer of 1990, NAGB retained an expert in standard-setting to assist its staff in conducting the item judgment procedure and analyzing its results. NAGB arranged for a panel of teachers, university experts, business leaders, and citizens to meet to perform item judgments in the fall. NAGB also retained a team of experts in testing, measurement, and evaluation to conduct a technical and policy evaluation of its procedures.

The item judgment task proved more difficult than expected. The results from the initial panel meeting and from a follow-up meeting later in the fall were inconsistent and were set aside early in 1991 on the advice of NAGB's evaluation team and other technical experts. These experts noted problems in NAGB's implementation of Angoff procedures and commented that the NAEP mathematics test included few items that represented "advanced" content and, thus, provided a weak basis for measuring advanced achievement.

NAGB formed new panels to apply its approach with modified procedures during the spring of 1991. The new panels produced more consistent results than before. NAGB's technical consultant reported that the achievement levels appeared technically defensible. However, NAGB's evaluation team noted that questions concerning NAEP's ability to measure advanced achievement remained, that there were still problems with NAGB's procedures and with the quality of the resulting data, and that the validity of NAGB's standards had not yet been examined. They

recommended that the results not be presented as standards and that the approach not be used further until it had been thoroughly reviewed.

NAGB judged the results sufficiently sound to be usable. It arranged to publish the results under its own authority in September and made them available to the National Education Goals Panel, which reported at the same time. Because NAGB's results were not issued as a NAEP report, they did not undergo NCES technical quality review.

NAGB also designed a request for proposals to apply the approach to the assessments of mathematics, reading, and writing for 1992. The 1992 contract, for $1.3 million, was awarded to an experienced testing firm, American College Testing (ACT). ACT has applied the NAGB approach with procedural changes that are discussed in chapter 2.

In their final report of August 1991, members of NAGB's evaluation team reiterated their concerns about the technical quality of NAGB's approach and its results. In its policy evaluation, the team observed that NAGB, whose members for the most part are not technical experts, may not be appropriately constituted to direct a testing program that must meet high standards of technical quality.

Objectives, Scope, and Methodology

In the fall of 1991, the chairmen of the House Committee on Education and Labor and the Subcommittee on Elementary, Secondary, and Vocational Education asked us to review NAGB's controversial exercise of its achievement goals responsibility. They asked us to answer three questions:

1. What are the strengths and weaknesses of NAGB's standard-setting approach?

2. Is NAGB's approach suited for use with NAEP, and might alternative approaches provide better benchmarks for goal achievement?

3. Are NAGB's knowledge resources and procedures sufficient to ensure that work done at its direction and the products that result are technically sound and that published measures meet appropriate standards?

Evaluating NAGB's Approach

To give ourselves a basis for answering the first question, we traced the legislative background of NAGB's achievement goals responsibility and

reviewed documents that recorded the development of NAGB's standard-setting approach. We examined the minutes of NAGB meetings from January 1989 to November 1991 and reviewed background papers, records of committee meetings, transcripts of public hearings, written testimony, technical memoranda and reports, correspondence, and NAGB and NAEP publications. We reviewed documents that described the NAEP mathematics assessment for 1990 and the procedures through which NAEP scale scores are estimated. In addition, we interviewed officials at NAGB, NCES, and ETS; NAGB's technical consultant and members of NAGB's evaluation team; and members of another evaluation team that reviewed the NAGB levels in connection with the Trial State Assessment (TSA).[11] Finally, we spoke with individuals who provided NAGB with written comments on the approach or who spoke at public hearings on the subject.

We drew our evaluation criteria from the standards issued by professional associations concerned with educational tests and measurement.[12] To understand the criteria that are likely to apply in a system of assessments linked to national content standards, we attended NCEST meetings and reviewed NCEST background papers as well as the final report. In addition, we examined the technical literature on methods of standard setting and on the application of the Angoff item judgment method. Our analysis compared NAGB's procedures against these general and specific criteria, identified novel or unconventional aspects of NAGB's approach and estimated their likely effects on test score selection and interpretation, used data from NAGB's technical reports to test for these effects, and drew conclusions regarding the technical soundness of NAGB's procedures. In addition, we checked the reasonableness of NAGB's 1990 results by locating other indicators of student achievement and comparing them to NAGB's findings.

We reviewed plans for the application of NAGB's approach to the 1992 assessments and preliminary and final results from that process for the mathematics assessment. (Work on the reading and writing standards was just beginning when we finished data collection for the draft of this report in June 1992.)

[11]The TSA is the state-level NAEP assessment of 8th grade mathematics undertaken on a trial basis in 1990. NAGB reported state results as well as national results in terms of the achievement levels. Accordingly, the NAGB levels were examined as part of the congressionally mandated TSA evaluation.

[12]American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, Standards for Educational and Psychological Testing (Washington, D.C.: 1985). Similar criteria are reflected in the Code of Fair Testing Practices in Education issued in 1988 by the Joint Committee on Testing Practices.

Evaluating Standard-Setting Methods Appropriate to NAEP

To answer the second question, we first drew conclusions regarding the appropriateness of NAGB's approach based on the analysis described above. We identified assessments of various kinds that measure student performance and apply standards (with an emphasis on national assessments) and interviewed experts involved with these possible alternative approaches. To evaluate the strengths and weaknesses of each approach for use with NAEP, we drew on our earlier work assessing the quality and use of NAEP data.[13] We also drew on our understanding of NAEP test design and scaling procedures and general requirements for technically sound description and trend reporting. We discussed the technical feasibility of possible alternatives with NCES technical staff.

Evaluating NAGB's Technical Capacity

To answer the third question, on whether NAGB's resources are adequate to support technically sound decisions, we examined NAGB's use of technical information in connection with the standard-setting project, drawing on the records and interviews described above. We asked officials of NAGB and NCES to identify additional cases that illustrate NAGB's handling of technical matters and reviewed records and interviewed participants in two such cases. To identify procedures that govern NAGB's decisions, we examined NAGB's written policies and a memorandum of understanding between NAGB and the Department of Education.

Study Limitations and Strengths

This study is based on published data that summarize the judgments made by participants in NAGB's standard-setting process in the spring of 1991 and focuses primarily on the 8th grade, the only grade for which detailed data on actual student performance were available. We checked the pattern of item judgments for the 4th grade also and found that it was consistent with our analysis. We did not check the 12th grade item judgments.

Our analysis of how students at different NAEP score levels perform on items of varying degrees of difficulty represents a new way of displaying what students can do and comparing it to what they should be able to do. Because this approach is new, its results should be considered suggestive rather than definitive.[14]

[13]U.S. General Accounting Office, Education Information: Changes in Funds and Priorities Have Affected Production and Quality, GAO/PEMD-88-4 (Washington, D.C.: November 1987).

[14]NCES commented that our approach leads to conclusions regarding actual versus expected performance that appear puzzling in light of other kinds of evidence. It has requested its technical review panel to conduct studies on the topic of comparisons.

A full examination of the strengths and weaknesses of alternative methods for setting goals or standards through NAEP tests was beyond the scope of this study. Our study is useful as a methodological critique of NAGB's approach, but because we could not be exhaustive in our research, we also cannot present definitive prescriptions without further research, analysis, and comparison.

The chief strength of this study is that we examined each step of NAGB's approach in relation to the others. Other evaluations have addressed specific aspects of the approach in greater depth than this study can offer, but none has examined the entire process and assessed its overall consistency. The results of our effort clearly show the importance of conducting a full step-by-step analysis prior to adopting a standard-setting procedure or accepting its results.

Agency Comments

Responsible officials from the Department of Education and from NAGB commented orally on a draft of the interim report that we completed in March 1992.[15] NCES officials generally concurred on the draft report. In oral comments and in correspondence after the report was issued, however, both the chairman of NAGB and the Assistant Secretary for Educational Research and Improvement argued that they believed the levels were appropriate and useful, that we had applied overly narrow technical criteria to what is essentially a judgmental process, and that we did not sufficiently credit improvements in procedures that NAGB implemented for 1992. Since NAGB selected its NAEP score standards on the basis of technical procedures, and since these scores were intended to be valid measures of the achievement levels that NAGB had defined, we consider technical criteria of evaluation appropriate. We take improvements in NAGB's 1992 procedures into account in chapter 2.

Officials from the department (for NCES) and from NAGB reviewed and commented on a draft of this report as well. The department's comments and our response to them are included in appendix I, and those from NAGB appear with our response in appendix II.

Organization of the Report

Chapter 2 of this report answers the first of the study questions by evaluating NAGB's approach as applied to the 1990 NAEP assessment and as revised for use with the 1992 assessment. Chapter 3, responding to the

[15]U.S. General Accounting Office, National Assessment Technical Quality, GAO/PEMD-92-22R (Washington, D.C.: March 1992).

second question, examines how NAGB's approach, modifications of that approach, and alternative methods might be used to set performance standards for use with NAEP. Chapter 4 examines the technical quality of NAGB decisionmaking, and factors contributing to technical quality, in response to the third question. Each chapter ends with recommendations.

Evaluation of NAGB's Approach

We were asked to evaluate the strengths and weaknesses of NAGB's approach to setting achievement standards and measuring how well students reached them on the 1990 NAEP mathematics test. NAGB's measurement effort included defining three achievement levels, using item judgment procedures to select a NAEP score for each level, and interpreting these NAEP scores in terms of the achievement level definitions and of statements of what students should know and be able to do.

Our evaluation began by examining the concepts of basic, proficient, and advanced achievement as NAGB defined them. NAGB's definitions are judgmental standards of what students at each level should know and be able to do. The definitions both guide the selection of a NAEP score for each level and provide the basis for interpreting student performance. Thus, it is critical to determine how well the concepts NAGB has defined correspond to what is measured by the NAEP scale.

Next, we focused on NAGB's item judgment and score selection procedures, asking whether the practices NAGB followed in setting its 1990 standards provided an adequate basis for judging how students at each level could be expected to perform on the 1990 NAEP mathematics test. We then used test score data and external evidence of student achievement to examine whether the scores NAGB selected can validly be interpreted in terms of NAGB's definitions and descriptions of what students at each level should know and be able to do. Finally, we reviewed how NAGB presented its results and whether it informed users of data limitations and cautioned against potentially unwarranted interpretations of the NAEP scores, as is appropriate whenever a new measure is introduced.

This chapter presents our findings on each of these matters with respect to 1990. Since NAGB changed its approach when it began further level-setting work in 1992, we also reviewed the changes to see if any problems found earlier were remedied. We conclude with recommendations for further action by NAGB and NCES.

NAGB's Achievement Level Definitions as the Basis for NAEP Score Selection and Interpretation

Our first step in evaluating NAGB's approach was to examine how NAGB defined the levels of achievement: to see what NAGB intended the achievement levels to represent. We then reviewed the NAEP scale, examined the achievement level definitions in light of what the NAEP scale measures, and identified problems that might arise in finding NAEP scores to match NAGB's definitions or, conversely, in using the definitions to interpret performance at various NAEP scores.

NAGB's Definitions

NAGB's definitions (summarized in table 2.1) are complex. They incorporate three aspects or dimensions of student achievement. First, the definitions refer to overall performance—to how much of the NAEP 4th, 8th, or 12th grade mathematics test a student can answer. Second, the definitions specify which types of items (what kind of material) the students at different levels are expected to master. Finally, NAGB's definitions link the achievement levels to predicted readiness for some future activity such as entry-level college coursework or advanced technical training. The standard-setting task was to find NAEP scores that would represent appropriate overall performance for each level, appropriate mastery, and appropriate readiness as well. This objective poses a challenge for the NAEP scale, which is designed simply to represent overall performance.

Table 2.1: Dimensions of Achievement in NAGB's Levels Definitions

	Dimension of achievement		
NAGB level	Overall performance on the NAEP test	Mastery expectation	Predicted readiness
Basic	Below proficient	Partial mastery of fundamental skills; for 12th grade, refers to standard high school work	Not specified
Proficient	Represents solid academic performance for the grade	Basic-level mastery, plus competency over challenging subject matter for the grade	Well prepared for the next level of schooling or (in the 12th grade) for adult life, work, and citizenship
Advanced	Superior performance, beyond proficient	Not specified	For 12th grade, ready for rigorous college courses, advanced technical training, or employment

Source: National Assessment Governing Board, The Levels of Mathematics Achievement, vol. 1, National and State Summaries (Washington, D.C.: 1991), p. 5.

NAEP Scores and Overall Performance

NAEP scores are a measure of overall performance on a test that covers a wide range of material, from very easy items that nearly 100 percent of students are able to answer to items that are so difficult that very few are able to answer them. The scores reflect both the number and the difficulty of the items answered correctly. They summarize how much of the test a student can answer, giving more weight to difficult than to easy items. (For more information on the NAEP test and scale, see appendix III.) Thus,

the NAEP scale can appropriately be used to set standards for how much performance (weighted by difficulty) is enough: what we call overall performance standards. (We discuss this type of standard in detail in chapter 3.)

NAEP Scores and Expected Mastery

The addition of expectations concerning student mastery, however, adds complications. Here, the question of interest is how students perform on certain types of items: for example, how students at the basic level perform on the items that test fundamental skills. NAGB's approach places considerable emphasis on identifying what students at each level should know and be able to do and on finding a NAEP score to match. But the NAEP scale measures—and NAGB's approach sets standards for—performance on the test as a whole, not performance on any particular type of items or class of skills.[1] Using the NAEP scale to express standards of what students could do as well as how much they should be able to do, we thought, raised important issues of procedure and interpretation.

The procedural issue was how to include all items in the item judgment process (as was necessary in order to use the NAEP scale) yet arrive at a standard that reflected mastery of the items that represented content pertinent to each level. The corresponding issue of interpretation was whether it was valid to interpret the attainment of the NAEP score selected for each level as evidence that the skills and knowledge expected at that level have been mastered (and to infer that lower scores mean that these skills have not been mastered). There is some question whether mastery inferences can be drawn from the NAEP scale under any circumstances.[2] In addition, we were uncertain whether NAGB's procedure, which uses the percentage of items answered correctly to summarize the performance expected at each level, would locate an appropriate score for each level on the NAEP scale.

NAEP Scores and Predicted Readiness

The definitions' predictions of readiness raise further issues of test coverage and valid interpretation. First, NAEP tests are not intended to be used for prediction. They were not designed with occupational skills or advanced college course prerequisites in mind—the 1990 mathematics test

[1]If content mastery were the sole dimension of interest, the most obvious measurement approach would be to identify items from the test that are relevant to each level, determine how many of those items students should be able to answer correctly, and find the proportion of students who did so. We discuss the application of content-based performance standards to NAEP in chapter 3.

[2]See Robert A. Forsyth, "The NAEP Proficiency Scales: Do They Yield Valid Criterion-Referenced Interpretations?" Educational Measurement: Issues and Practice, 10 (1991), 3-9 and 16.

did not cover calculus, for example—so coverage of these skills or prerequisites (unlike coverage of grade-level material) cannot be assumed.

Second, predictive standards cannot be based on judgment alone: they must be backed by factual information. Establishing a valid predictive standard generally requires identifying what skills and knowledge, at what level of difficulty, someone must master in order to be prepared for future success and making sure that the standard reflects performance on these relevant items. The alternative is to show that, content coverage aside, scores on the test actually predict success as claimed: for example, to demonstrate that, for whatever reason, a 1990 NAEP score of 330 (12th grade proficient) is the dividing line between the 12th graders who succeed in freshman mathematics courses or on the job and those who do not. (We return to this point in our discussion of predictive validity.)

Summary

All in all, NAGB's definitions imply three different measurement purposes that are difficult to conciliate and raise fundamental questions about whether the achievement levels can be adequately measured by the NAEP test and scale or can be validly used to interpret performance at various test scores. We looked at NAGB's 1990 procedures (summarized in table 1.1) with these questions in mind.

The Adequacy of NAGB's Score Selection Procedures

We examined NAGB's 1990 item judgment procedures to determine whether panelists were given the resources they needed to do their job. These resources include a clear understanding of what students at each level should be able to do with respect to the material covered on the test and a basis for making informed judgments of how students at different levels are likely to perform on the various test items. Next, we reviewed the steps by which the item judgment results were transformed into a score on the NAEP scale.

NAGB's Item Judgment Procedures

NAGB gave panelists its definitions of basic, proficient, and advanced achievement and instructed them to make prescriptive judgments—to state how students marginally qualified for each level should perform on every item on the 1990 NAEP mathematics test—using these general definitions as a guide. We concluded that this part of the approach had two important weaknesses.

First, NAGB did not provide or ask panelists to develop a clear statement of what students at each level should know and be able to do with respect to 4th, 8th, and 12th grade mathematics before starting the detailed work on items (as is commonly recommended in the literature for the application of Angoff item judgment methods). Instead, NAGB left it to each panelist to formulate and apply his or her own working definition or standard of what students at each level and grade should know and be able to do.[3] The absence of common prior standards makes it difficult to trace the connection between NAGB's definitions and panelists' judgments and leaves the item judgment results themselves as the only basis for inferring the skills and knowledge expected of students at each level and at the corresponding NAEP score.

The second difficulty with NAGB's procedure was that panelists' judgments were not backed by sufficient information—information that would have helped them understand how students marginally qualified for each level would be likely to perform on different types of items. Most importantly, panelists were not assisted in reaching informed judgments concerning how students functioning at the basic level might actually perform on difficult items—items that such students should not (by definition) be expected to answer and would most likely guess at or leave unanswered. What percent correct judgment should be given to these difficult items? Inappropriately high percents correct would push the overall score higher with the result that it would represent performance beyond what is required in the definition.

Prior experience with item judgment methods suggests that the solution to this problem is to have judges estimate how students marginally qualified for a given level will do on a question and help them by giving them data on how such students (students who meet the standard in question) actually perform on very difficult questions. If judgments are realistic in this sense, hard questions that do not belong in the expectations for a fundamental-skills level will be given a realistically low percent correct estimate and thus will have appropriately little effect on the overall score standard.

NAGB did not, however, direct panelists to make their judgments realistic or provide the data needed to do so. Panelists did have data on the percentage of all students who answered each item on the 1990 NAEP mathematics test correctly, but those are not helpful in estimating how

[3]NAGB has recognized that specification of the achievement levels in terms of mathematics was a problem in 1990. In its 1992 procedure, panel members developed common working definitions specific to the subject tested before beginning their item judgment work.

basic-level students, for example, perform.[4] Also, panelists were not assisted in understanding what percentage of students could get an item right simply by guessing.[5] In the interest of setting ambitious standards, panelists may have resolved any uncertainties by pushing up the percents correct, and NAGB's procedures afforded little protection against such unrealistically high judgments. The problem here is that item judgments may have gone beyond the expectations stated in the definitions for the basic and possibly for the proficient level.

In summary, the item judges did not start their work with a common framework identifying the mathematics skills and knowledge appropriate to each level and had little basis for judging how students at each level would most likely answer difficult NAEP items. These gaps in NAGB's procedures left only one basis for the item judgments: individual panelists' views. The question arises whether these individual views constituted a reliable basis for setting a national percent correct standard for each achievement level.

The Issue of Reliability

A standard based on item judgments is said to be reliable if there is evidence that (1) individual panelists were consistent in the judgments they made and (2) the views represented among the panel were reasonably representative of those found among qualified judges generally, such that the average of panelists' ratings is a trustworthy estimate of a more general standard. The reliability of the standard (its freedom from measurement error because of the composition of a particular panel) is typically assessed by examining the degree of variability in the judgments expressed.

The design of NAGB's 1990 levels-setting procedure did not permit a full examination of the reliability of the judgment data. However, it was possible to examine the consistency of means across the four regional panels whose judgments formed the basis for the standards. NAGB's technical report on the 1990 achievement levels acknowledges that there was "substantial and troublesome variability" in estimates of the basic level of achievement across these four panels. This variability could have stemmed from real regional differences in standards, but it could also have been the result of the procedural weaknesses we have discussed. (NAGB's

[4]NAGB provided panelists at the first item judgment meeting with visual displays that show how the probability of getting an item correct changes as NAEP scores increase, but these displays were too technical to be useful. Subsequent panels in 1990 and 1992 received only overall performance data.

[5]Panelists apparently made their own individual estimates of the results of student guessing. NAGB recognized that this might be a problem, especially if the panelists' methods were different from those used in forming the NAEP scale, but did not see how it could be solved.

report recommended that the sources of measurement error be evaluated in future applications of the approach.)

Whether panelists' judgments were well informed or not, NAGB's procedures needed to identify the point on the NAEP scale that matched those judgments. We now examine how this was done.

Transforming Item Judgment Results Into a NAEP Score

To transform the item judgments into a NAEP score, NAGB summarized the item judgments by calculating the percent correct for each level across all items and then found a matching percent correct point on the NAEP scale. To illustrate, item judgment panelists set 48 percent correct as the standard for the 8th grade basic level, and 48 percent corresponds to a score of 255 on the NAEP scale. All the scores chosen as standards, we found, represented the percent correct specified by the judges. However, we questioned whether these scores represented the performance expected for each achievement level for two reasons:

1. The percent correct summarizes how many items students should be able to answer, without regard to whether the items are easy or difficult. Thus, it loses information about the expected pattern of performance on items of varying difficulty—information that is important to the distinction between the achievement levels.[6]

2. The NAEP scale is not a simple measure of the percent of items answered correctly. Rather, the scale score reflects the pattern of performance on easier and more difficult items. To illustrate, 48-percent correct achieved by answering only the easier items (as might be expected of a basic-level student) corresponds to a lower NAEP score than getting 48 percent correct by answering both easy and difficult items. NAGB's procedure finds a central point in this range of scores that represents 48-percent correct, but this central point may not represent the pattern of performance appropriate to the basic-level student.

Taking these two problems into account, the question arises whether NAGB's procedures identified the point on the NAEP scale at which performance indeed matched the expectations for different levels of

[6]The rationale for using the percent correct mechanism is strongest when the test is designed to measure performance at the standard being judged—that is, when most of the items on the test are items that most students who have just reached the standard (but not those who fall below the standard) will be able to answer. If a test contains large proportions of items that are either too easy or too difficult with respect to the standard, the rationale for using the percent correct is weakened considerably.

mastery of different types of items expressed in the item judgments. If not, it would be invalid to interpret the NAEP scores in terms of these expectations as NAGB's approach seeks to do.

Summary

We found that NAGB's procedures did not resolve the issues of measurement and interpretation raised by its definitions of the achievement levels. To the contrary, our review of NAGB's procedures reinforced our concern about these issues. We concluded that it was important to look closely at the test scores that NAGB's approach selected to see whether the inferences of overall performance, mastery, and readiness implied by NAGB's definitions and achievement level descriptions were in fact supported.

Validity of Interpretation: What Do NAGB's NAEP Score Standards Represent?

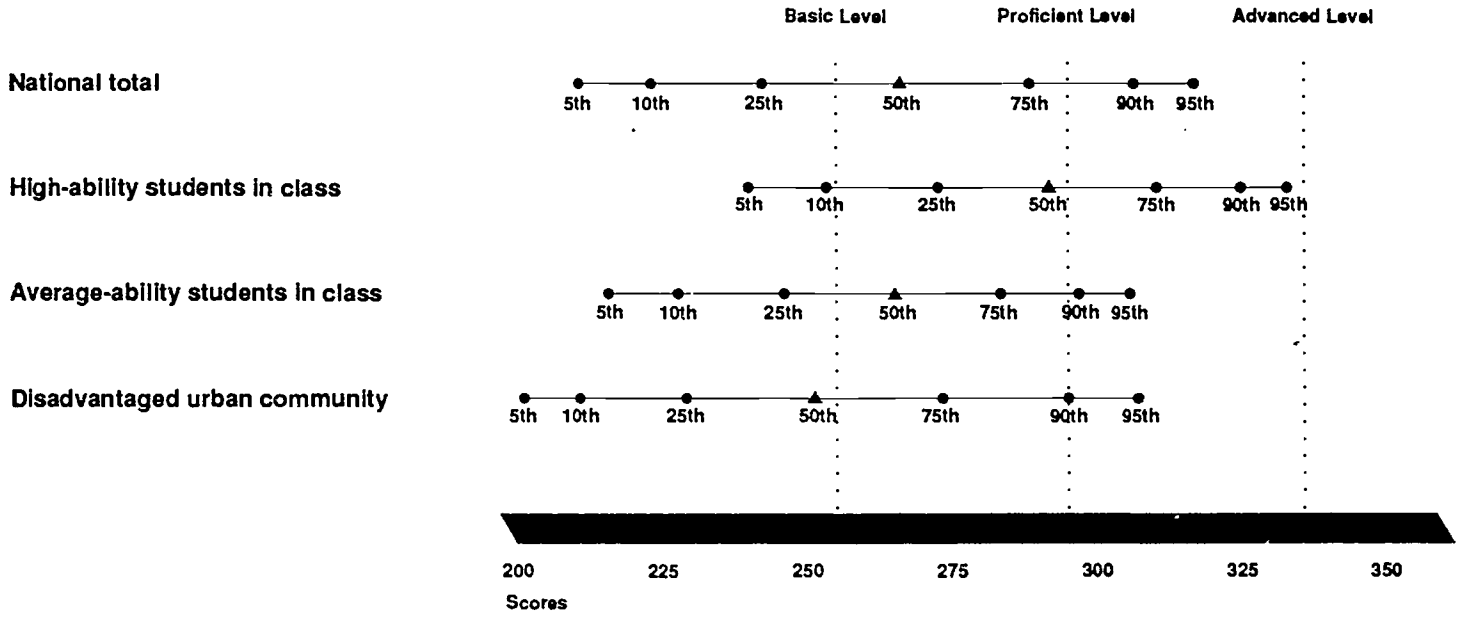
Our validity review examined whether NAGB's results and the interpretations given to them were consistent with other indicators of basic, proficient, and advanced achievement. We found that the literature on item judgments stresses the importance of conducting such a review as part of the judgmental process of setting standards. NAGB's achievement levels policy paper recommended that validity studies be conducted, but no such studies were undertaken for 1990.[7] To conduct a preliminary evaluation of validity, we checked NAGB's results against several readily available indicators of overall performance, mastery, and readiness.

NAGB's Standards as Indicators of Overall Performance

The most straightforward interpretation of NAGB's results is that the score for the advanced level represents superior performance, that the score for the proficient level represents solid performance for the grade, and that the score for the basic level represents performance that is something less than solid. A glance at how American students performed on the 1990 NAEP mathematics test (figure 2.1) confirms that the score for the advanced level represents superior overall performance, one that few students even in high-ability classes were able to reach. The score for the proficient level represents above-average performance even for students in these classes. Clearly, performance at this level can reasonably be said to be solid, and the score for the basic level is substantially less than this, as by definition it should be.

[7]NAGB titled its spring 1991 item judgment procedures plan a "replication/validation study." However, this plan simply provided for new item judgment panels; it did not include any assessment of the validity of the interpretations given to the standards produced by these panels.

Figure 2.1: NAGB Achievement Levels and NAEP Score Distributions (Percentiles) **in 8th Grade Mathematics, 1990**



We also checked NAGB's results against data on the mathematics achievement of 9-year-olds and 13-year-olds from the 1990-91 International Assessment of Educational Progress (IAEP). The IAEP test was similar to NAEP, although the content was adjusted to be reasonably representative of curricula across the various participating nations. We identified the score attained by the top 1 percent of U.S. 9-year-olds on the international test (equivalent to the percentile of 4th graders that NAGB's approach classified as advanced) and identified the proportion of students from other nations who equalled or exceeded this score. The same procedure was applied to the scores of 8th graders and 13-year-olds. (The IAEP report did not provide sufficient detail to allow similar comparisons at the basic and proficient levels.)

Fewer than 5 percent of the 9-year-old students in any nation tested demonstrated advanced achievement according to this comparison. For 13-year-olds, 10 percent of the students in Taiwan and at least 5 percent of those in China (restricted sample) met this standard; in no other nation did as many as 5 percent meet the advanced threshold. This comparison

indicates that NAGB's advanced level represents superior performance even by world standards.

The question is, What more do these scores represent? Is it valid to infer that they represent the points at which students have attained the skills and readiness that NAGB's definitions and level descriptions imply? For example, is it valid to infer that nearly 40 percent of all 8th graders, over half of those in disadvantaged urban communities, and 10 percent of those in classes that teachers identified as containing "high-ability" students have not even partially mastered fundamental skills as figure 2.1 and NAGB's definition of the basic level would suggest? To examine these questions, we first examined how students at the scores NAGB selected for each level actually performed.

The Mastery Interpretation of NAGB's Standards

To examine whether student performance at the NAEP scores NAGB selected in 1990 matched the mastery expected for each level, we calculated the percentage of students that judges thought should answer items of varying difficulty and compared it to the answers that students at the NAEP score for each level actually gave. First, using the final round of 1990 item judgment data, we divided the 137 questions on the 1990 8th grade mathematics test into four groups—the 35 easiest items, the 34 moderately easy items, the 34 moderately difficult items, and the 34 most difficult items—and then calculated the percentage of students who should answer items in each group correctly at each achievement level. This gave us the pattern of performance on different groups of items that judges expected at each level.[8] (For further information on the performance patterns, see appendix VI.)

Next, we estimated the percentage of students at each level (that is, at the NAEP scores of 255, 295, and 336) who actually did answer items in each group correctly.[9] We then examined whether student performance at each NAEP score matched panelists' expectations with respect to items that they should have mastered (that is, items that 80 percent or nearly 80 percent of students should have been able to answer).

[8]We used judgment data from the final 1990 round of item judgments, and we used 8th grade as our example because actual percent correct data by achievement level were available for this grade but not for the others. We calculated the performance patterns for the 4th grade based on item judgments also and found them to be very similar to the 8th grade patterns.

[9]We used actual percent correct data for the 61 items made public from the 1990 test. There were at least 13 such items in each difficulty group, and the percent correct figures by item group for these items were comparable to those for the entire item set. We divided the items into groups based on data showing the percentage of all students (out of those who attempted an answer) who actually got each of the 137 items correct.

Our results are shown in table 2.2. With respect to items that should have been mastered, the comparison shows that at the basic-level NAEP score of 255, students did considerably better than they should have according to NAGB's judgments on the items relevant to that level. Students at t' : proficient-level score of 295 also performed better than they should have on relevant items, although the difference is not great. Students at these levels fell below panelists' expectations regarding performance on items that they were not expected to master. For example, only 16 percent of basic-level students answered the most difficult items correctly, compared to the item judgment expectations of 26 percent. For the proficient levels, the figures were 38 percent compared to 56 percent. These data suggest that judgments with respect to these items were unrealistically high and that their inclusion pushed up the standards for these levels.

Table 2.2: Panelists' Judgments Compared to Actual Performance on Items Relevant to Each Achievement Level

	Percent correct	
Achievement level and item group	**Item judgment expectations**	**1990 test results**
Basic level (255)		
Easy Items	67%	83%
Moderately easy items	50	60
Proficient level (295)		
Easy items	85	93
Moderately easy items	75	84
Moderately difficult items	67	67
Advanced level (336)		
Easy items	96	96
Moderately easy items	92	95
Moderately difficult items	88	90
Most difficult items	80	71

Source: National Assessment Governing Board, The Levels of Mathematics Achievement (Washington, D.C.: 1991), vol. 3, Technical Report, pp. 265-71 (item judgment data), and vol. 2, State Results for Released Items, pp. 3-35 (percent correct data).

This evidence suggests that while students at the basic and proficient score standards can be said to have met the mastery expectations for their level, students at lower scores may have matched these expectations as well. For the basic level, the matching score could have been considerably lower (around 240-42); at this score, the basic standard would have been met by a considerably larger percentage of students than NAGB reported (75 percent rather than 62 percent).

In contrast, students at the advanced level did less well than they should have according to NAGB's judges on the most difficult items. However, they achieved the required overall percent correct by performing better than expected—indeed, nearly flawlessly—on the easier but less relevant three quarters of the test. This again suggests caution against invalid inference: overall performance on a test that consists largely of standard grade-level materials does not necessarily imply mastery of items that are "advanced" in the sense that they require complex thinking.

The Predictive Validity of the 12th Grade Standards

The NAGB definition of proficient achievement at the 12th grade level states that students at or above that level are prepared for college study and for productive work. We examined information relevant to these predictions.[10]

We found that NAGB's paragraph describing what proficient 12th graders should know (see appendix V) was generally consistent with the College Board's summary of what students need to know to undertake postsecondary study in mathematics.[11] As noted in chapter 1, NAGB found that only 16 percent of 12th graders had met or exceeded the proficient level and thus should be considered ready for college work. We calculated that this 16 percent represents about 27 percent of the students who enroll directly in higher education after high school.[12] If only 27 percent of freshmen are prepared for college study in mathematics, then 73 percent are not prepared. The American Council on Education reported, in contrast, that 29 percent of incoming freshmen in 1991 identified themselves as needing remedial work in mathematics.[13] Even allowing for the fact that not all students may recognize their need for remedial work, the gap again raises questions about interpreting the achievement level scores in terms of NAGB's definitions.

To evaluate the reasonableness of the proficient standard with respect to preparation for productive work, we compared NAGB's description of 12th

[10]The definition also predicts readiness for democratic citizenship and for responsible adulthood. We did not examine these predictions.

[11]College Board, Academic Preparation for College (New York, N.Y.: 1983).

[12]The Bureau of the Census has estimated that 60 percent of 1990 high school graduates enrolled in higher education the following fall. (This figure is reported in U.S. Department of Education, National Center for Education Statistics, The Condition of Education, 1992 (Washington D.C.: 1992), p. 28.) We assume that virtually all the students who score in the top 16 percent on NAEP are part of the college-bound group. Sixteen percent of the total 12th grade population is equivalent to 27 percent (16/60) of the college-bound population.

[13]A. W. Astin et al., The American Freshman: National Norms for Fall 1991 (Los Angeles: Higher Education Research Institute, UCLA, 1991), p. 14.

grade proficient performance to findings from the Secretary's Commission on Achieving Necessary Skills (SCANS Commission), to findings from a 1990 New York State field study of the mathematics skills requirements of a wide range of jobs that do not require a 4-year college degree, and to the skills covered in certification tests for six such occupations.[14] Our review suggested that the skills NAGB described in connection with the basic level would be sufficient for productive employment in most jobs that do not require a college degree. The materials we reviewed indicate that while many such jobs require the application of modest levels of knowledge of arithmetic, measurement, and probability, very few use even the simplest algebra or geometry.[15]

With respect to readiness for rigorous college study, we compared NAGB's finding that 2.6 percent of 12th grade students reached the advanced level with the report that 1.5 percent of U.S. 11th and 12th graders took advanced placement exams in calculus in 1991 and 1 percent scored well enough to be deemed eligible to receive college credit. This suggests that NAGB's standard identifies a percentage of the student population roughly comparable to the percentage that are not only ready for college study but have already undertaken it while in high school. The 1990 NAEP test did not cover calculus, and there is no way of knowing whether the two procedures identify the same segment of the student population.

Conclusion: Validity of NAGB's Score Interpretations

We conclude that the NAEP scores selected through NAGB's procedures are incomplete and somewhat misleading representations of the achievement levels. Student performance at the NAEP scores selected cannot validly be interpreted in terms of NAGB's definitions or of the item judgments. The NAEP scores cannot be used to find the percentage of students who have met the content mastery and readiness criteria NAGB defined for each level.

Flaws in procedure, such as the lack of information support to panelists (especially of level-specific performance data) and possibly the use of the percent correct mechanism to translate item judgment expectations into a NAEP score, contributed to the lack of correspondence between the scores

[14]U.S. Department of Labor, Secretary's Commission on Achieving Necessary Skills, Skills and Tasks for Jobs (Washington, D.C.: U.S. Government Printing Office, 1992), and New York State Education Department, "Report to the Board of Regents on Career Preparation Validation Study," Albany, N.Y., n.d. We reviewed the occupational tests administered by ETS for cosmetology, production and inventory control, cash management, construction code inspection, construction supervision, and food protection inspection.

[15]The argument can be made that the high-performance workplace of the future will require workers to have more extensive quantitative knowledge than do most current jobs. But NAGB's readiness criteria are stated in terms of readiness for any productive work. The proficient level is supposed to represent what all students need, not just what students need to enter a field that relies on mathematical skills.

NAGB selected and the interpretations offered for them. Failure to validate whether the NAEP scores selected matched the expectations for mastery expressed in the item judgments or in the level definitions constituted a further important flaw.

Most important of all, and contributing to the fundamental measurement problems inherent in NAGB's approach, is the critical conceptual issue raised at the beginning of this chapter—the issue of whether the mastery dimension of NAGB's achievement levels can adequately be represented using the NAEP test and scale, which were designed to depict overall performance. Whereas the procedural flaws suggest merely that NAGB's approach was poorly executed, the conceptual issue speaks to the approach itself: our data suggest that it was invalid for the purpose of drawing inferences about content mastery. We will return to this issue in chapter 3.

NAGB's Presentation of the Results

NAGB's aim in establishing achievement levels was to make NAEP scores more interpretable—to enable readers to compare current performance against a standard of what students should know and be able to do. As we have seen, however, NAGB's interpretations of the achievement levels (which confound what panelists thought students should do with what they actually could do) have been misleading.

NAGB described its procedure as the application of standards that represent broad consensus on what students should know, which is also misleading. This language creates the impression that the achievement level descriptions represent general content mastery standards—standards that were developed independently of, and subsequently applied to, NAEP data.[16] In fact, the descriptions were a product of (not a guide to) the standard-setting process; they represent test items that met certain narrow selection criteria.

NAGB surveyed various user groups to ascertain whether they found the achievement levels useful, and most did. Many user representatives responded that the levels report had clearly conveyed the significance of student performance, but others found the report unclear. We note that

[16]For example, on the basis of the description of the levels in the National Goals Report, officials of the U.S. Metric Association took NAGB's levels descriptions to be general standards and to imply that only advanced 8th grade students are expected to understand metric measurements. They expressed concern to GAO. When we explained the very restricted bases of the levels descriptions, these officials commented that referring to these descriptions as standards of what students should know was very misleading.

respondents did not have information that would have alerted them to weaknesses in NAGB's approach and data, nor were efforts made to explain and rule out competing plausible explanations of NAGB's findings of poor student performance. Indeed, the survey did not ascertain whether respondents had interpreted the data accurately.

NAGB was aware of some of the limitations of the levels data when it published the 1990 levels results in September 1991 but neither cautioned readers concerning the reliability of the data nor noted that validity had not been established. The technical report that appeared late in November did offer cautions on reliability grounds. NAGB itself was unaware of other significant limitations, largely because a thorough review of its basic approach had not been conducted. (Such a review is required in preparation of all NAEP reports, which are published through NCES, but not for reports published under NAGB's independent authority. NCES officials have told us that the 1990 levels results probably would not have passed this review.)

Moreover, although NAGB's achievement levels report did describe its 1990 standard-setting as a trial effort that was based on a process that "while imperfect, was serviceable," it presented the results of that process as "revealing and diagnostic" and urged those seeking to make change to take action based on NAGB's standards and data.[17] Such advice was premature.

The 1992 Levels Procedures

NAGB was aware of some of the flaws in its procedures and acted to correct them for 1992. It secured the services of a contractor, ACT, that is experienced in standard-setting and in the use of item judgment methods and has access to considerable technical expertise. ACT has proposed and implemented improvements that have strengthened the item judgment procedure. These include careful attention to panelist selection, improved training, the development of guiding definitions for each subject and grade prior to beginning the item judgment process, and review of the reliability of the judgment results. However, critical weaknesses inherent in NAGB's overall approach remain unaddressed, as indicated in table 2.3.

[17]National Assessment Governing Board, The Levels of Mathematics Achievement, vol. 1, National and State Summaries (Washington, D.C.: 1991), pp. 11 and ix.

Table 2.3: 1990 Weaknesses and 1992 Procedures

1990 weakness	1992 procedure
1. Levels definitions incorporate dimensions ill-suited to the NAEP scale; no clear standard for performance in mathematics	Original definitions not changed; panelists developed a working standard for each level and grade by applying the definition to the NAEP mathematics test framework
2. Standard for each level is based on both relevant and irrelevant items	Unchanged
3. No level-specific performance data to enable panelists to estimate realistically	Unchanged
4. Reliability of item judgments not fully assessed	Reliability being examined; results not yet available
5. Percent correct score used to summarize item ratings, assumed to capture pattern	Unchanged. Use of a pattern-based method has been found to change scores only slightly
6. No examination of validity of inferences made from NAEP scores	The question of inference being examined as part of the NCES prepublication review process

In addition, the 1992 procedures pose new issues and problems. The working definitions created by the item judgment panels (item 1 in table 2.3) provide the specific standards for performance in mathematics that were missing from the 1990 approach. NAGB plans to use these working definitions of what students should be able to do to interpret student performance at each NAEP score selected as a standard. These include a working definition that makes clear that the advanced level represents complex understanding of mathematics, not simply stronger performance on a wide range of problems. As we have shown, however, NAGB's procedure cannot be assumed to select a NAEP score at which students' mastery matches judges' expectations. Unless a good match can be demonstrated, the working definitions cannot be used to interpret performance at the score selected.

Finally, NAGB's standard-setting procedure was applied to reading and writing for the first time in 1992. The reading and writing assessments made much more use of extended-response questions than the mathematics exam. The item rating procedure, which is easy to apply to multiple-choice questions, must be adapted for use with extended-response questions. ACT pilot-tested procedures for judging such items and concluded that they were feasible, but it is too early to say whether the actual judgment panels were successful. New procedures and data sources may be required to check validity for the reading and writing standards.

NAGB's publication policy specifies that the achievement levels were to be the basis for reporting NAEP results for 1992, and NAGB has directed NCES to instruct ETS to use the achievement level NAEP scores as benchmarks for reporting student achievement and to offer the judgment panelists' working definitions as interpretations of performance at each of these scores. NAGB's policy also specifies that all NAEP reports must meet NCES technical quality standards, which are similar to those applied in this report. In light of the many questions we and others have raised about NAGB's approach, the commissioner has several times urged that NAGB continue to use the old method of reporting NAEP results until the achievement level approach can be shown to be sound. On each occasion, NAGB has reaffirmed its commitment to the use of the achievement levels.

Conclusions: the 1992 Approach

We conclude that while ACT's 1992 procedures have addressed some of the problems that affected the 1990 standard-setting, the fundamental problem of finding a test score that can validly be interpreted in terms of NAGB's definitions and descriptions remains unaddressed. If anything, the gap between the level definitions, the achievement level descriptions, and actual performance at the NAEP score selected for each level is likely to be greater than before. Unless and until NAGB can show that its approach is internally consistent and produces valid interpretations of the NAEP scores selected to represent each level, it should either refrain from reporting in terms of the achievement levels at all or present the levels scores simply as NAGB's judgmental standards for partial, solid, and superior performance, without further interpretation.[18]

Overall Conclusion

As NAGB noted in its 1990 policy on setting achievement levels, its approach was an experiment. NAGB concluded that its experiment was sufficiently successful to warrant the continuation of the approach, with procedural improvements. However, it did not analyze the effects of the modifications it made in the Angoff process, nor did it examine whether the scores it selected could be validly interpreted in terms of the levels definitions. Having done so, we reach a different conclusion.

[18]The 1992 NAEP results for mathematics were published in April 1993. The achievement levels did undergo NCES prepublication review. NCES accepted NAGB's levels-setting process and the resulting scores as given and focused on the question of valid inference—on whether NAGB's statements of what students at these scores should do are useful for interpreting what such students actually can do. Early results from technical evaluations suggested a need for further examination of the inferences that can be drawn from NAEP scores. Both the achievement level approach and the conventional anchor point approach are now under review. The 1992 NAEP report presents data both ways, alerts readers to the questions that have been raised, and urges readers to assess for themselves how well the various forms of reporting meet their needs.

We conclude that NAGB's 1990 approach was inherently flawed, both conceptually and procedurally, and that the evaluation team's advice—that the approach not be used further until a thorough review could be completed—was warranted. NAGB's approach identified scores that represent different levels of overall performance, but these scores are not necessarily evidence that students have the skills and knowledge specified for that level. NAGB's approach was not changed in fundamental respects for 1992 and is likely to produce unsupported and quite possibly erroneous interpretations with respect to the 1992 NAEP tests also.

These weaknesses are not trivial; reliance on NAGB's results could have serious consequences. For example, policymakers might conclude that since nearly 40 percent of 8th grade students did not reach the basic level (a NAEP score of 255), resources should be reallocated so as to emphasize fundamental skills for most classes. Since many students who scored below 255 were in fact able to answer basic-level items (according to our analysis), this strategy could retard their progress toward mastering more challenging material. Similarly, parents or educators might conclude that NAGB's description of the advanced 4th grade level (which simply summarizes certain difficult test items) represents the mathematics skills that should be taught to gifted children and focus their curriculum accordingly. Other testing entities might adopt NAGB's procedures, on the understanding that they produce valid and useful results. And finally, NAEP might abandon its existing straightforward empirical basis for score interpretation in favor of one that is unrelated to actual performance.

Recommendations

In light of the many problems we found with NAGB's approach, we recommend that NAGB withdraw its direction to NCES that the 1992 NAEP results be published primarily in terms of levels. The conventional approach to score interpretation should be retained until an alternative has been shown to be sound.

Second, we recommend that the Chairman of NAGB and the Commissioner of Education Statistics develop a joint plan and schedule for a review of NAGB's achievement levels approach (its definitions of achievement, score selection procedures, and score interpretation), taking into account evaluations that are currently under way and providing for additional activities as needed. The plan should begin with a review of existing critiques of the approach and should include, at an early stage, a determination by the commissioner whether (1) NAGB's approach will necessarily produce invalid interpretations of NAEP scores and should not

be pursued or (2) the approach is sufficiently promising that a specific plan for preparing for NCES prepublication review should be designed and implemented.[19] If option 1 is selected, the case is closed. If the decision is to proceed, NAGB should develop evidence that the levels results are valid and reliable and that the interpretations suggested for them are supported. NCES should make clear what evidence will be required.

[19]As discussed in detail in chapter 4, the commissioner is responsible for ensuring that NAEP depicts achievement fairly and accurately and for conducting reviews and validation studies of the assessment.

Alternative Standard-Setting Approaches

The second study question asked us whether NAGB's approach is suited for use with NAEP and whether alternative approaches might provide better standards for goal achievement. In chapter 2, we concluded that NAGB's achievement levels approach, which sets standards for overall performance on NAEP but interprets them in terms of what students at each performance level should have mastered, is unworkable. In this chapter, we do not discuss NAGB's approach further. Instead, we consider the general question of how NAEP might be used to set goals, benchmarks, or standards. Our analysis distinguishes between overall performance standards and content-based performance standards. We begin this chapter with a general discussion of these two types of standards and then evaluate alternative approaches to each one. The chapter ends with recommendations.

Two Types of Performance Standards

Student performance standards can take either one of two forms. Overall performance standards identify how much performance is enough and are expressed in terms of a total score on a test of knowledge that is generally relevant to the standard (that is, a test that focuses on the material that students are expected to know at the levels of difficulty they are expected to be able to handle). Overall performance standards are typically used to determine whether a student knows enough to be considered qualified for some specific purpose such as high school graduation or professional certification. This type of standard may also be used to group scores according to descriptive categories such as unqualified, marginally qualified, and fully qualified. The standard is generally measured in terms of the number or proportion of questions answered correctly. Any pattern of right answers that yields a score at or above the overall performance score standard is acceptable, since all items on the test are deemed relevant to the standard.

Content-based performance standards, in contrast, indicate that a student has mastered specified subsets of content at an acceptable level. Only performance on items whose content is pertinent to the standard is taken into account in setting the standard and in measuring student performance in terms of the standard. If a test covers several content areas (such as algebra, geometry, and statistics), each of which is the subject of a standard, items are sorted by content area and a separate scale is formed and a separate standard is set for each area.

The two standards serve different purposes, provide different kinds of information, and require different properties of the test on which they are based. Differences between the two types of standards are summarized in table 3.1. Our analysis focuses on whether NAEP tests are adequate to support each of the potential alternative methods we have identified, how each method might be applied, and how its results can be interpreted.

Table 3.1: Overall Performance Standards and Content-Based Performance Standards

Overall performance standards	Content-based performance standards
Purpose: To group test scores into categories that represent levels of performance or qualification for some stated purpose. Key question: Have students learned enough?	Purpose: To identify the performance that signifies achievement of a content mastery standard. Key question: Have students learned what they should have learned?
Prerequisite: Agreement on what constitutes the standard. Standards may take the form of ordered descriptive categories such as acceptable and outstanding or may be defined in terms of the proficiency needed to perform successfully in some current or future status	Prerequisite: Content standards (statement of the knowledge and skills that students are expected to master)
Performance standard: Score on the test as a whole	Performance standard: Score based on items pertinent to each content standard
Test requirements: Test must cover the knowledge areas specified at an appropriate level of difficulty	Test requirements: Test must be aligned to the content standards and must contain sufficient items pertinent to each standard to support accurate measurement
Interpretation: Someone who scores at or above the performance standard is said to know enough to meet the standard	Interpretation: Someone who scores at or above each performance standard is said to have mastered the content addressed in that standard

The choice of method must ultimately be guided by the choice of purpose. If the purpose of setting national standards for student performance is to determine whether students have reached an acceptable overall level of mastery of grade-level material, overall performance standards are appropriate. Overall performance standards, however, do not indicate whether students have mastered the skills and knowledge specified in national content standards. Content-based performance standards serve that purpose.

Setting Overall Performance Standards for NAEP

NAEP tests are designed to cover knowledge generally relevant for a given subject and grade and, thus, provide a potentially suitable basis for applying standards of overall performance on grade-level material. Since NAEP tests are designed to be most accurate in the range of performance typical of the majority of students, they are potentially well suited for expressing overall performance standards applicable to the majority of students. As suggested in table 3.1, it is important to identify key knowledge areas and the level of difficulty that students should be expected to handle in each area and to design NAEP tests accordingly.[1]

NAEP can be used to set standards for levels of performance that represent a challenge to today's average student. However, NAEP is likely to provide insufficient data to support accurate measurement at very high levels of performance that very few students reach. Care must be taken to ensure that criteria for and interpretations of high overall performance are expressed in terms of consistent mastery of the various grade-level materials covered on the test, rather than mastery of the most difficult items specifically.

We examine three methods through which overall performance standards for NAEP might be set: methods based on current performance, on criterion performance, and on test items. We found that while these methods can be used singly, they are commonly used in combination.

Standards Based on Current Performance

Overall performance standards can be selected by examining how various categories of students currently perform on the test, interpreting the overall capabilities represented by test scores at various points on the scale, and putting these two kinds of information together to select scores to be used as benchmarks or standards for future years. The selection of scores under this first method is truly a matter of informed judgment, reached by consensus. The method consists of defining categories of overall performance, assembling individuals representing diverse views of what students can and should achieve on a NAEP test, giving them data on actual performance and assistance in interpreting performance at various score levels, and asking them to select a score or ranges of scores to represent each overall performance category.

[1]In the past, NAEP tests have concentrated on materials that are covered in most classrooms, in effect using the existing common curriculum to define what students should learn. The common curriculum, however, does not necessarily represent expert and citizen consensus concerning what should be taught. To measure progress with reference to emerging standards as well as to past and current practice (as is NAGB's policy for future tests) requires striking a delicate balance in test design.

In one variant of this approach, the basis for setting standards is performance-distribution data. Standard-setters examine the distribution of scores for the total sample and for subgroups of students and select total scores that appear to be suitable benchmarks for the various performance categories. (Categories can be given labels such as marginally qualified and fully qualified or can be expressed in terms of target percentiles—for example, as the score that at least 80 percent of students should achieve.) Patterns of item mastery associated with each score are examined to confirm that the score is appropriate in terms of the knowledge and skills expected for each category, and the potential benchmark is adjusted upward or downward until members of the panel reach consensus that a score that represents the performance category in question has been found. Examples based on NAEP might include selecting the score that represents the current 75th percentile for students in high-ability classes as the standard for "excellent" performance or using the 15th percentile score achieved by students in the top two thirds of schools as the 15th percentile standard for the nation as a whole. In a second variant, standards are set by identifying several test items that typify each level of proficiency to be established, determining the point on the overall performance scale at which most students are able to answer each typical item, and selecting a midpoint score or upper and lower boundary of each proficiency category based on these data.

In either variant, the selection of performance-based standards is a matter of prescriptive judgment supported by data. (Judgments based on overall performance might consult any of the types of data we used in chapter 2 in examining the validity of NAGB's score interpretations. Judgments based on item mastery might seek evidence that the benchmark items are good indicators of mathematics proficiency, relatively unaffected by factors such as reading proficiency or familiarity with a particular item format.) The chief requirements for setting standards based on inspection of current performance are that the selection body's claim to represent broad consensus be credible and that its decisions be supported by logic and information.

The methods just described can be used with existing NAEP tests and do not require that detailed performance criteria be spelled out in advance. Standards selected on the basis of actual performance (with attention to student diversity) send a message that what is accomplished in some schools should be expected of all. The chief drawbacks of performance-based methods are that judgments may be inadequately informed and (at the other extreme) that data-gathering may consume

undue time and effort and bury decisionmakers in details. These drawbacks can be prevented by agreeing in advance what key viewpoints and what types of data will be considered and how diverse views will be taken into account and by providing decisionmakers with analytic support.

Standards Based on Criterion Performance

Overall performance standards may also be set by drawing on independent information concerning students' mastery of the material covered by the test. The general steps in this method are

1. describe the characteristics of students at each performance category to be established—for example, what the marginally proficient, proficient, and exceptionally proficient student should be able to do with respect to the materials covered on the test;

2. find individual students or groups of students who are independently judged to match these descriptions;[2]

3. ascertain how these individuals or groups perform on the NAEP test and what NAEP scores would be estimated based on their performance; and

4. use these data to find a benchmark score or range of scores that represent qualified performance.

The General Educational Development (GED) testing program provides an example of this type of procedure. The GED examination is a multisubject test battery covering core knowledge in high school subjects and is used to determine whether a student who did not complete high school is qualified to receive a high school equivalency certificate. The program uses graduation from high school as its criterion. It regularly administers its tests to samples of graduating seniors and, using these results, adjusts the minimum passing score on the GED examination so that it reflects the 30th percentile: the score that divides the lower 30 percent of diploma holders from the upper 70 percent.[3] Similarly, the advanced placement

[2]It may also be useful to identify students who would definitely not be expected to meet the criterion, such as students enrolled in the next lower grade.

[3]The choice of 30th percentile represents a policy judgment by the GED program, which we present for illustrative purposes. Participating states are free to (and do) select higher figures, in line with expectations for and the performance of their own graduates.

program calibrates the scoring of advanced placement tests to the performance of students in college courses.[4]

For some types of standards, criterion groups cannot be readily located. For example, there is no simple way to find a group of "fourth graders who have mastered the skills fundamental to challenging fourth-grade work" (the appropriate group for NAGB's "basic" level of achievement). Where there is no obvious criterion group, experts can be asked to identify individuals or test papers that exemplify the performance associated with the criterion. For example, teachers whose classes were selected for inclusion in the NAEP sample might be trained to understand NAEP performance categories and asked to identify and to enter a code on the test booklets of students whose classroom performance meets the criteria for each category. Alternatively, experts involved in the grading process could be asked to identify completed test papers that exemplify performance that meets those criteria. NAEP scores projected on the bases of the exemplary papers could then be examined.

We conclude that criterion performance methods could be used with current NAEP tests. Criterion performance methods might require that tests be given to groups not included in the normal sample or that new procedures be instituted to identify exemplary papers within that sample.

Criterion performance methods have an advantage in that they build in attention to empirical validity. The corresponding disadvantage of such methods is that they work well only when criterion performers can be readily identified. Given the imprecision of all tests, scores associated with criterion performers are likely to be somewhat spread out, and it may be difficult to find a single score that distinguishes qualified or exemplary students from others. Since NAEP scores are not being used to make decisions about individuals, however, there is no requirement that a benchmark selected to represent a category of performance be set at the bottom of the achievement range associated with that category. A benchmark that represents a central point in that range may be equally useful.

Standards Based on Test Items

In a third approach, overall performance standards can be established through judgments concerning performance on test items, as in the standard Angoff method. The province of Alberta, Canada, for example,

[4]Decisions as to whether to grant college credit are made by colleges, which can and do set standards that exceed those suggested by the testing program.

identifies standards for excellent performance on provincial assessments for elementary school subjects (expected of the top 15 percent of students) and adequate performance (expected of at least 85 percent of students) by applying item judgment methods to the test as a whole.[5]

We find that NAEP tests could form the basis for applying overall performance standards through Angoff item judgment procedures, provided that the standards were not aimed at extremes of achievement (where NAEP is less accurate) and that the criteria for sound implementation of judgment procedures were met. Criteria for sound implementation of item judgments include (1) clear, consensus-based specification of what students at each level should be able to do on the full range of items on the test; (2) panelists who have the expertise and information to make accurate judgments of how students who actually have those capabilities will perform on each test item; and (3) review of student performance at the NAEP score selected to confirm that it matches panelists' expectations.

Concluding Observations on Overall Performance Standards

The methods discussed above incorporate different sources of information about student performance and thus suggest different bases for interpreting the test scores selected as standards. Whatever the method or combination of methods for setting the standard (actual performance, criterion performance, or item judgment), it is important to confirm and to be able to provide evidence to validate or support any interpretations that are given. Whether the item judgment process, which is expensive, contributes sufficient interpretive information—beyond what can be obtained through the other methods—to be worth the cost is a question that merits further review.

Feasibility of Overall Performance Standards: Current Studies

Our review suggests that, in principle, each of the methods discussed could be applied to current or future NAEP tests, as long as the performance standards are expressed in terms of levels of mastery of the materials that make up the test as a whole. So far, only one method of standard-setting has been applied to NAEP. Experiments with additional methods, however, are in progress. Both ACT and the National Academy of Education (NAE) are currently comparing NAGB levels results to actual

[5]Teachers make the item judgments, guided by statements of expectations based on the curriculum and by their own experience of what students in the upper and lower 15 percent of the distribution can do.

performance data.[6] NAE also proposes to have experts construct achievement level scores based on items independently judged to be appropriate to each achievement level.

These efforts will provide valuable information about the feasibility and usefulness of alternative standard-setting methods, as well as revealing whether these methods yield results that are consistent with NAGB's. When their results are in, NAGB and NCES will be in a good position to compare the feasibility of and types of information produced by alternative methods of setting standards or by combinations of approaches.

Setting Content-Based Performance Standards Through NAEP

As shown in chapter 2, current NAEP tests and the NAEP scale were not designed to support content-based performance standards. (NAEP's purpose and design are summarized in appendix III.) However, existing tests might contain sufficient items to support some content-based performance standards, and future tests might be designed to match such standards. A general difficulty with the application of content standards to NAEP is that the United States has not established national standards describing what students should learn, nor is there a national curriculum. Thus, the usual prerequisites for the identification of content-based performance standards are not in place.[7] National content standards, however, are on the way. Standards are currently being developed for mathematics, science, history, and geography—subjects regularly assessed by NAEP—under conditions that are intended to produce the expert and citizen consensus essential to widespread adoption.

We examined how existing NAEP tests might address content standards once these have been developed and adopted and how such standards might be accommodated within NAEP's design. Our analysis assumes that monitoring (that is, measuring current student performance and tracking changes in performance over time) will continue to be the major purpose of the assessment and that any changes designed to align NAEP to content standards will need to be compatible with that purpose. We further assume that NAEP will be part of a larger system of standards-related

[6]The NAE study is part of an evaluation of the NAGB levels under contract to NCES.

[7]We have analyzed experiences in the Canadian provinces with developing tests and standards in connection with a predetermined curriculum. See U.S. General Accounting Office, Educational Testing: The Canadian Experience With Standards, Examinations, and Assessments, GAO-PEMD-93-11 (Washington, D.C.: April 1993).

assessments of student performance at the classroom, school, district, state, and national levels.[8]

Applying Standards to Existing Tests

Whether and how an existing NAEP test can be used to measure student performance in terms of a particular content standard will depend on whether the test contains a sufficient pool of items appropriate to that standard. NCES officials estimate that at least 20 items are needed to measure a skill or content area accurately, and the items need to be reasonably representative of the domain of content described in the standard. Where there is an adequate pool of items to sustain accurate measurement, a content-based performance standard might be set for these items through item judgments or any other appropriate technique.

For example, suppose that national mathematics standards define "core" proficiency for 8th grade in content terms. Experts might be convened to identify the NAEP items pertinent to core proficiency and to evaluate whether the item pool was sufficiently large and representative of the required skills. If there proved to be enough items that they could be legitimately combined to form a separate subscale, item judgment procedures (or indeed any of the procedures discussed in connection with overall performance standards) might be used to suggest the score that represents adequate mastery of the material. Or NAEP might be used to set a standard of adequate performance in mathematics with respect to each component area covered by the test (see appendix III) using the subscales that were incorporated into the test by design.

Where there are not sufficient items to support a content-based performance standard, NAEP can report the percentages of students who answered illustrative items correctly and can identify the overall NAEP scale score at which items illustrative of the standard are consistently mastered. These strategies do not provide a measure of achievement of a standard, but they provide useful interpretive information.

The chief advantage of taking NAEP's current design as "given" is that this strategy ensures that NAEP will continue to serve its statutory purpose of supporting fair and accurate description and trend analysis. The chief disadvantage is that some content standards—those not suited to

[8]NAGB solicited public comment in the fall of 1992 regarding the role of NAEP in a system of assessments. NAGB's discussion paper asks readers to consider whether national content and performance standards should determine the content of each assessment or whether content should continue to reflect (evolving) current curriculum as well as these standards. It also asks for views on the continued use of the achievement levels and whether other approaches to identifying appropriate achievement goals should be considered.

measurement through NAEP—may be left unassessed at the national level. For example, a standard that every student should develop specialized knowledge of one out of six areas of applied mathematics would be difficult to measure through NAEP, because adding enough items on all six areas to ensure that each student tested can show knowledge of one such area would make the test unworkably long.

Designing NAEP to Fit Content Standards

We find that, in principle, NAEP could be designed to address national content standards but that technical considerations limit the range and type of content standards for which NAEP can provide performance measures. Within a design that is intended to describe overall performance accurately, based on a sample of students each of whom sees only a portion of the test questions, NAEP will be best able to address standards that concern commonly taught materials that most students are expected to master and least able to address standards that concern specialized materials or knowledge to which few students are currently exposed.[9]

The implication is that there are significant limits to what can be expected of NAEP in measuring student performance against content standards. Assessment of selected areas of content as well as different levels of performance in each area is practically impossible because there will never be room in the test for all the necessary questions. NAGB hoped to do both; the more realistic choice is one or the other.

That is, NAEP could potentially be used to set a standard for general mastery of each of several broad skill areas (such as the component areas included in the 1990 mathematics exam—numbers and operations, measurement, data analysis, algebra and functions, and geometry) but not multiple levels of mastery within each area. Alternatively, NAEP could be designed to include items in various content areas at several levels of complexity and to measure the percentage of students who have mastered each level of complexity. By virtue of its design, NAEP cannot provide data pertinent to standards of either kind that apply only to small groups of students, such as those in specialized programs. Content mastery standards that refer to skills that cannot validly be assessed through a

[9]Students (and especially less-proficient students) tend not to answer questions about material that they have not had the opportunity to learn. The greater the proportion of unanswered questions, the less accurately NAEP can estimate student performance. A substantial increase in the proportion of items that represent unfamiliar material—especially if it is material to which only advantaged or very able students are exposed—may thus decrease NAEP's ability to describe the performance of average and below-average students accurately.

brief paper-and-pencil test (such as the skill of carrying out a long-term project independently) will also fall outside NAEP's scope.

In summary, designing NAEP to fit national standards of general student mastery of commonly covered content and establishing standards for student performance with respect to these content standards could improve the usefulness of the assessment. However, designing NAEP to support standards that refer to specialized or emerging content areas raises a host of technical and policy questions and is also clearly premature, given that no standards have yet been developed or adopted. Until more is known about the nature of the standards for different subject areas and how these will be addressed by other elements in the assessment system, it is difficult to envision the test design and analysis techniques that would be most appropriate or how NAEP could best contribute to the overall system.

NAGB and NCES could usefully explore how NAEP might respond to different types of standards within and outside of the traditional sample and design. For example, NAEP might use the traditional sample to address only the standards that refer to the mainstream curriculum but create experimental modules for use in a subsample of schools or in states whose curricula include these practices and report the results separately. Discussion and pilot-testing of possibly useful techniques now could help NAEP prepare for the day when standards become available.

Conclusions

We conclude that while NAGB's particular approach is not suitable for use with NAEP, alternative approaches are feasible. Approaches designed to set standards of overall performance seem more readily applicable to NAEP than content-based standards. Although they require consensus on relevant skills and skill levels, overall performance standards do not presuppose the existence of national content standards and do not require changes in the design of NAEP tests. Within the next few months, NAGB and NCES are likely to have data sufficient to evaluate whether any of the methods, or a combination, could result in standards that would be accepted as credible on both technical and policy grounds. However, overall performance standards cannot be used to assess whether students have mastered particular content at an acceptable level.

Approaches designed to measure content mastery —other than a simple item benchmarking approach—would be more difficult to apply and could require changes in test design. Content mastery is a matter of considerable

concern, and NAEP will undoubtedly be expected to monitor progress toward the achievement of national content standards. Therefore, we conclude that activities to explore how NAEP can be designed to support content-based standards without compromising its ability to serve its statutory purposes are also needed.

Overall performance standards and content-based performance standards provide different information and represent different kinds of achievement goals. Our analysis suggests that it is important to clarify which of these kinds of achievement goals NAGB and NAEP should address.

Recommendations

In view of the conceptual and technical flaws inherent in NAGB's achievement levels approach (see chapter 2) and of the many questions that need to be resolved before an alternative standard-setting method can be selected, we recommend that NAGB withdraw its policy of applying the 1990 achievement levels approach to future NAEP tests and join with NCES in exploring alternatives for setting both content-based and overall performance standards with respect to NAEP. This inquiry should examine issues of purpose, technical feasibility, cost, fairness, credibility, and usefulness.

We recommend that the Congress specify what it intends in directing NAGB to identify appropriate achievement goals: whether it envisions the establishment of overall performance standards, the establishment of content-based performance standards, or simply better alignment of test coverage with content mastery standards. Given that legislation to establish a mechanism for adopting national content standards is currently under consideration, the Congress may also wish to express specific guidance with respect to activities to align NAEP to content standards before such a mechanism is in place.

The Technical Quality of NAGB's Decisions

Our final study question concerns whether NAGB, whose members for the most part are not technical experts, has knowledge resources and procedures sufficient to ensure that work done at its direction will be technically sound. We examined NAGB's actions in planning and implementing the achievement levels approach with this question in mind. We also examined two other technical decision areas, so as not to base conclusions on what could have been an atypical example.[1]

Our evaluation of the achievement levels example takes account of the fact that NAGB's approach to this highly important task was new and unusual. In addition to making untested modifications to Angoff item judgment procedures, NAGB reversed the usual sequence of steps involved in testing student performance in terms of standards. NAGB started with a test and derived standards of what students at three levels should know from the test items, using the item judgment process. The usual sequence is to develop broad expert and citizen consensus on what students should know, determine whether one or more levels of achievement should be measured, and design a test or tests accordingly. As illustrated by Canada's approach to national standards and testing and by the work of the National Board for Professional Teaching Standards in the United States, preparing for and developing standards-based testing takes much time and requires extensive technical support.[2] We looked for evidence that NAGB considered the feasibility of its quite different approach carefully at the planning stage. Bearing in mind that difficulties with a new procedure cannot always be foreseen, we looked especially for evidence that NAGB reviewed the results of its 1990 approach carefully before concluding that it was sound.

In all three of the examples we reviewed, we looked for evidence that the technical feasibility and implications of the issue under discussion were examined and alternative approaches were evaluated, where feasible and appropriate, before action was taken; that technical procedures were adequately planned and supported; that expert advice was sought at key points during policy planning and implementation; that issues raised by such experts were addressed and resolved; and that the products of technical procedures were reviewed according to appropriate quality criteria before they were accepted.

[1]We did not review NAGB's handling of other kinds of decisions, such as decisions regarding the subject matter to be covered in each assessment. Our findings with respect to technical issues do not reflect on NAGB's performance in other areas.

[2]On the Canadian approach, see U.S. General Accounting Office, Educational Testing: The Canadian Experience With Standards, Examinations, and Assessments, GAO/PEMD-93-11 (Washington, D.C.: April 1993).

NAGB has indicated that it considered the identification of achievement goals to be a matter of informed judgment rather than a technical issue, and it believes that it is not appropriate to apply criteria of technical quality to the setting of achievement levels standards. We agree that standard-setting is a matter of informed judgment. However, when the selection of test scores to serve as standards rests wholly or largely upon technical procedures and data, as it did in the achievement levels case, it is not only appropriate but also essential to inquire whether those procedures and data were technically sound and based on adequate information. Whatever the method by which scores were selected, it is important to verify that the interpretations given to those scores are supported by evidence.

The Achievement Levels Setting Case

We presented a descriptive summary of the background, development, and implementation of NAGB's approach to setting performance standards in the form of achievement levels in chapter 1. For the present chapter, we analyze NAGB's use of technical information at each key point in these events.

Planning and Design

Our analysis suggests that NAGB adopted and modified the Angoff item judgment method as the basis for its approach without sufficiently examining its requirements and limitations and, indeed, may have misunderstood what this method can produce. The Angoff method was first recommended to NAGB in a December 1989 staff paper, which proposed that Angoff judgments be used to set a standard of performance on items representing core knowledge for each grade, a purpose for which this method is well suited. However, the paper stated that the Angoff method is used to identify "core" items through panelists' yes-no judgments and that the method would identify NAEP scores that reflect expected performance on these core items only. As we have indicated in earlier chapters of this report, the Angoff method used with the NAEP scale does not set a standard for performance on core items: it sets a standard for performance on the test as a whole. The statement in the NAGB paper is not only incorrect but seriously misleading.

NAGB committees subsequently recommended the use of Angoff procedures based on the understanding that this is the most practical of the item judgment methods and can be applied to an existing test. However, our search of staff papers, committee records, and transcripts of NAGB meetings revealed no detailed examination of the limitations and

requirements of the Angoff method and provided evidence that existing alternative methods received only a cursory review.[3]

We also found little evidence that NAGB examined the technical implications of its policy decision to set multiple standards using the entire NAEP test as the item base or that it understood that the resulting scores would not necessarily represent mastery of "core" items for each level.[4] From what we can determine, NAGB simply assumed that having diverse panelists judge every test item with respect to all three achievement levels, guided by generic definitions of the three levels, would produce reliable results, although the technical literature suggests otherwise.

Although the December 1989 concept paper and the May 1990 policy paper were circulated for comment, NAGB did not formally develop a technical design for the initial item judgment process and did not obtain a technical critique of the overall design before going ahead. (NAGB did get reviews of the materials developed to orient judges but not of the overall plan. Moreover, the schedule did not allow for procedures to be pilot-tested.) When the initial design proved problematic, NAGB's staff designed revised procedures. The redesign was only sketchily developed and reviewed before it was implemented, and it introduced changes that further departed from standard Angoff practice and made it difficult to evaluate the reliability of the resulting data.

Use of Expert Advice: 1990

We found that NAGB did solicit and receive sound technical advice. Most of the methodological and procedural issues that proved problematic were first raised during the period in which the policy was under consideration, many of them by NAGB members and staff. Much good advice regarding improvements to the item judgment procedures was followed, in 1990 and for 1992. However, NAGB set aside experts' early advice to proceed cautiously and examine alternative methods more fully before selecting a standard-setting approach. Perhaps more important, NAGB did not respond

[3]NAGB considered using traditional NAEP anchor points or the scores for the 25th, 50th, and 75th percentile of students as benchmarks. We found no discussion of other methods of setting performance standards such as those we presented in chapter 3.

[4]NAGB's May 1990 achievement levels policy paper states that the item judgment panels would use "a proven judgment procedure to recommend which test questions and/or which proportion of questions students need to answer correctly to reach different achievement levels." (National Assessment Governing Board, The Levels of Mathematics Achievement, vol. 3, Technical Report (Washington, D.C.: 1991), p. 345.) In fact, the procedure identifies only the proportion of questions to be answered correctly. Transcripts of the March and May meetings make clear that some NAGB members did not understand the Angoff method or thought that they did not have enough information about it.

to fundamental questions about the NAGB approach that emerged from the evaluation of its initial results. Although NAGB initially emphasized that its approach was provisional, it has not opened it to reconsideration despite recommendations from several quarters that it do so.We find that NAGB approved the 1990 achievement levels results and extended its commitment to the levels approach without adequate evidence that its procedures and results were technically sound and led to valid interpretations of NAEP scores. (Our criteria for evidence of technical soundness and validity were presented in chapter 2 of this report.) NAGB's May 1990 paper noted that the 1990 levels-setting is a trial but stated NAGB's expectations that its results would be usable for 1992 and that the levels would be the primary basis for reporting in 1992 and beyond. In November 1990, NAGB was informed of technical difficulties with the levels approach but nonetheless adopted a publication policy that stated that the achievement levels should be the primary basis for reporting NAEP results and took action to incorporate achievement levels into the request for proposals for 1994-96 NAEP operations.

NAGB approved the 1990 results for publication and solicited the contract proposals for the application of its levels approach to the 1992 NAEP assessments in mathematics, reading, and writing before the reliability of the levels results had been fully evaluated, in the absence of any confirmation of the validity of the interpretations given for them, and against the advice of its evaluation team. It instructed NCES to set up the 1992 NAEP results in terms of the achievement levels and reaffirmed these instructions as recently as August 1992, although preliminary review of the mathematics results for 1992 (including the statements of what students at each level should be able to do) had indicated that the approach was still possibly problematic.

Our analysis of the record suggests that several factors contributed to NAGB's decisions to move forward with the levels approach despite the questions that had been raised about it. As already noted, NAGB considered the selection of test score standards to be a matter of policy judgment and did not recognize the degree to which validity of interpretation would be an issue. NAGB recognized that score selection should be based on a defensible procedure and had been advised that although questions about it remained, the 1990 item judgment process met this criterion. The procedures used to construct the achievement level descriptions and illustrative items, which were based in part on standard NAEP methods, appeared defensible as well. The benefits of sending an important message about U.S. students' school achievement appeared considerable, and NAGB

saw little risk in publishing scores and interpretations that had yet to be fully examined. NAGB's idea was that these results could be validated and adjusted later if necessary.[5]

Changes for 1992

NAGB recognized that it had underestimated the technical resources needed to implement the item judgment procedures and therefore greatly increased the resources available for 1992. Rather than being done in-house, the 1992 item judgment procedures were conducted through a $1.34 million contract to ACT, a testing firm with extensive experience in standard-setting and an expert staff, aided by advice from external experts. The budget and timelines provided for analysis of the reliability of the item judgment results. However, the budget was sufficient to fund only relatively small judgment panels, which may limit the reliability of the results. A more significant limitation is that by specifying that the 1992 contractor should implement its 1990 approach, NAGB probably discouraged bidders from proposing a technically stronger design.

Most significant of all, the contract for the 1992 standard-setting did not provide for validity studies to be undertaken. NAGB has now remedied this omission and has entered into a separate contract with ACT for this purpose.

Decisions concerning publication of the 1992 NAEP mathematics results had to be made before these studies could be completed. The NCES prepublication review did not address the levels scores or the procedure by which they had been reached: these were taken as given. Rather, the review focused on the issue of valid inference—on whether NAGB's statements and illustrations of what students at each level should be able to do were a valid basis for understanding what students at those levels actually can do. Questions arose about the achievement levels and about the traditional anchor points as well. The report summarized these questions, indicated that studies to resolve them were under way, presented test results in terms of both methods, and urged readers to compare them and to join in the debate on how to reflect standards through NAEP reporting.

Conclusion and Implications

We conclude that in the case of the achievement levels, NAGB undertook a technically complex function that it lacked the specialized knowledge to

[5]In fact, new levels scores for 1990 were calculated as a result of the 1992 standard-setting process. The descriptions of the skills and competencies associated with each level were also revised.

direct. Unfortunately, the technical nature of this function was not evident at first. NAGB viewed the selection of achievement goals as a question of social judgment that NAGB, by virtue of its broad membership base, was well suited to decide. The Angoff method and the interpretation of the scores selected through that method did not appear to pose technical problems—as indeed they might not have, if NAGB had followed its original plan to set a single standard for each grade. The decision to set three standards per grade (although it was well supported on policy grounds) created a host of technical complications that NAGB neither recognized nor addressed. The result of this chain of events is that NAGB has given policy direction to NCES to take actions concerning NAEP that are technically unsound.

The implications of the achievement levels example are significant. NAGB's authorizing statute assigns it several functions that are explicitly technical in nature, such as developing test specifications and the methodology of the assessment. Other functions, such as the development of assessment objectives, are not explicitly technical but must be performed with an awareness of NAEP's technical capabilities and limitations and of the requirements for accurate estimation and trend reporting. The example of the achievement levels raises concerns that, in the absence of sufficient knowledge on technical issues, NAGB could give directions that might undermine NAEP's technical integrity and render it unable to serve its statutory purposes.

However, the achievement levels example may be atypical. Adopting the levels policy was one of NAGB's earliest major actions, undertaken at a time when working relations between NAGB and NCES had not been fully established. Setting performance standards was new for NAEP, so there was little direct experience to guide NAGB's deliberations. Users such as the National Education Goals Panel were eager to see NAEP results reported in terms of standards, for use in the panel's first report on the achievement of national goals. The technical deficiencies we observed in the process might simply reflect these specific circumstances.

Other NAGB Actions in Technical Areas

To see whether NAGB's actions in the levels case were typical of its technical decisions generally, we reviewed two additional cases. In one case, NAGB undertook to identify the conditions under which states or other testing entities would be permitted to link their own assessments to NAEP and to set up procedures for evaluating and responding to proposals of this nature. Requests from state testing directors prompted action in

this case. The issue was clearly technical, and NAGB recognized this from the first. It arranged for experts to review the policy options and drafted general guidelines for handling requests to link other assessments to NAEP, in which technical review was delegated to an NAGB committee and to NCES. After hearing concerns from the Council of Chief State School Officers, NAGB arranged for further technical and constituency review so that the policy could be refined to fit varied circumstances.

The issue in the other case was whether to continue to report NAEP results in terms of a single scale that covers all three grades tested (cross-grade scaling) or to report each grade in terms of its own scale (within-grade scaling). This issue raises a question of policy: should NAEP be designed to provide detailed information about variations in performance within a given grade, or should it sacrifice some within-grade detail in order to show increases in learning from one tested grade to the next? To answer this question, however, requires an understanding of the kinds of information that various types of scaling can provide. It also requires attention to the cumulative or grade-specific nature of the subject areas tested. Thus, sound policy choice depends on an understanding of technical issues.

In this case, NAGB initially focused on the policy question and on the advantages each type of scale has in principle. Responses from NCES and ETS, among others, suggested that a specific review of the current cross-grade scaling and of the effects of changing to within-grade scaling for NAEP should be undertaken before a decision was made. NAGB subsequently obtained such a review through NCES's technical review panel. This review supported NAGB's initial preference for within-grade scaling but also found that NAEP's cross-grade scaling is well implemented. A member of NAGB who is also a curriculum expert argued strongly that topics cannot be neatly divided by grade level in some subject areas and that options should be left open. NAGB adopted a policy that calls for the use of within-grade scaling where feasible and appropriate. Each subject area consensus group will be asked to determine which type of scale best fits its needs. NCES has noted that these groups, like NAGB itself, contain few members whose training would allow them to grasp the technical implications of this question.

Both of these cases proceeded differently from the achievement levels effort. NAGB recognized that each case involved technical issues on which it should seek expert advice. NACB arranged for technical information to be analyzed and for issues and alternatives to be fully examined by technical

experts. In both cases, NAGB postponed action on a draft policy when an important constituency asked for further review. In both, NAGB abandoned an initial "one size fits all" approach for a policy that could be adapted to varied circumstances. Finally, NAGB's actions and policy decisions in both cases made use of the technical resources available through NCES and were consistent with the expert advice that NAGB received.

These three cases indicate that as might be expected of a body designed to represent a broad spectrum of constituency opinion, NAGB has tended to focus initially on the policy dimensions of the issues that have come before it. Its handling of the technical dimensions has varied. In one case, NAGB immediately recognized the technical aspect of the policy issue (linking state tests to NAEP), sought expert review of the options, adopted initial decision guidelines, and delegated implementation to technical experts in NCES. In the second, NAGB recognized immediately that within-grade scaling was a technical issue as well as a policy issue, provided for in-depth technical analysis after learning of NCES and ETS concerns, but delegated decisions to groups that include only limited expertise in scaling. In the third case, NAGB conceived of standard-setting as a policy function that it should itself perform, did not adequately examine whether the approach it designed was technically sound, and set aside technical experts' concerns.

We conclude that NAGB has access to significant technical resources through NCES and through its own ability to contract for expert services. When NAGB recognizes the technical dimension of a policy area, it can use these resources appropriately to inform policy planning and to implement policy guidelines. However, NAGB may not recognize that a policy has important technical dimensions or may subordinate technical quality to policy requirements. In such cases, NAGB could unknowingly (or unintentionally) give policy direction to NAEP that is not technically sound.

Contributing Factors

In search of explanations for the problems evident in the achievement levels case and for strategies that might ensure that NAGB follow the useful practices observed in the two other cases, we examined three factors: the structure, responsibilities, and capabilities of NAGB and of NCES, NAGB membership, and NAGB operating procedures.

Structure, Responsibilities, and Capabilities

We examined NAEP's governance structure to determine the protection it offers against the receipt of technically unsound policy direction. We also

considered whether such protection as does exist is likely to be effective and looked for structural and procedural features that foster policy direction that is both technically sound and responsive.

By statute, NAEP is governed through a structure of two units, each with a unique strength. (Appendix IV summarizes the statutory structure.)

One unit, NAGB, is designed to be broadly representative. NAGB is a lay body composed of members of key constituencies (state and local officials and educators, citizens, and two experts in measurement) who meet several times a year with committee activities between meetings; it is assisted by a staff of six professionals. NAGB is independent of the Department of Education. It has a general responsibility to formulate policy guidelines for NAEP and to advise the Commissioner of Education Statistics (who heads NCES) regarding the conduct of the assessment. The Secretary of Education (through the commissioner) reports to NAGB regarding actions to implement NAGB's decisions.

NAGB is responsible for carrying out specific functions or responsibilities, which it may delegate to its staff. These include selecting the subject areas to be addressed; identifying appropriate achievement goals; developing assessment objectives; developing test specifications; designing the methodology of the assessment; developing guidelines and standards for analysis, reporting, and dissemination; developing standards and procedures for interstate, regional, and national comparisons; and taking action to improve the form and use of the assessment. NAGB has the final authority on the appropriateness of cognitive items, is directed to ensure that items are free from bias, and directs the consensus process through which test objectives (the content areas to be covered on each test) are established.

The other unit, NCES, is characterized by technical expertise. NCES is staffed by full-time technical experts and has access to others through its advisory committees. It has access to additional experts through ETS as the technical contractor for NAEP and through the technical advisory body that ETS is required to consult. NCES is responsible for administering NAEP and for ensuring that the assessment is fair and accurate. NCES is also responsible for conducting review and validation studies of NAEP and has established a technical review panel for this purpose.

The Commissioner of Education Statistics, who heads NCES, is the guardian of the quality of statistical data produced under his or her

supervision, including NAEP data. NCES reviews all statistical data prior to their publication, using standards established in consultation with the Associate Commissioner for Statistical Standards and Methodology (by law, a highly trained expert) and the Advisory Council on Education Statistics.

The potential for conflict in this structure—including the possibility that NAGB might give NCES policy direction that the commissioner, as guardian of technical quality, could not implement—has long been evident.[6] Designed to ensure that policy guidance for NAEP is free of inappropriate influences, it gives NAGB responsibility for many functions that are highly technical but does little to ensure that NAGB's judgments are technically well informed. The structure makes NCES responsible for NAEP's technical quality, but NCES's primary technical quality control mechanism, the prepublication review process, comes into play only after a policy has been implemented and produced results. Moreover, the NCES review process does not apply to reports published under NAGB's independent authority. (As already noted, this was the path followed for NAGB's 1990 achievement levels publications.)

NAGB Membership

Our review suggests that NAGB needs to have enough technically trained individuals among its members to ensure that the technical implications of policy issues are recognized early and are given appropriate attention throughout the policy planning process. Its formal structure provides for two members who are experts in testing and measurement. During the period covered by our review, however, only one of these two had strong technical training; the other person's expertise lay in assessment policy and implementation.

We conclude that it is not sufficient for NAGB to have only one member with strong technical training. A single individual's efforts may be spread too thin; his or her absence from a particular meeting will leave an important perspective unrepresented; and a single spokesperson for an unfamiliar view is unlikely to prevail in a group discussion. The two positions specified in the law are a bare minimum for adequate decisionmaking and should probably be increased.

[6]NAGB urged the Congress to rationalize NAEP's structure in December 1989 and again in January 1991.

Operating Procedures

We located two sources of procedural protection against technically unsound policy direction: the memorandum of understanding between NAGB and the Department of Education signed in April 1992 and NAGB's operating policies.

The Memorandum of Understanding

The memorandum of understanding was negotiated in an attempt to resolve the conflicts and ambiguities in NAEP's governance structure. The memorandum commits the department to "make every reasonable attempt to implement the policy-setting actions" of NAGB and specifies that when such actions cannot be accomplished, the department and NAGB will seek mutually satisfactory resolutions. This specification suggests that if NAGB's direction conflicts with the commissioner's statutory quality-assurance responsibilities, the commissioner may legitimately inform NAGB that he or she cannot follow its instructions and may refuse to accept as satisfactory a solution that is not technically sound.

We found that while this memorandum appears to give the commissioner the right to act in accordance with his or her judgment and statutory responsibilities if given technically unsound direction, it does not fully resolve the issue. Judging from the continued negotiation over whether 1992 NAEP results should or should not be reported in terms of achievement levels, it is not yet clear what constitutes a "reasonable attempt" or what will happen if a mutually satisfactory solution cannot be found. We were assured by NAGB's executive director and by its members that NAGB can envision circumstances under which the commissioner could legitimately be unable to follow NAGB's directions. NCES officials, however, appeared unconvinced that NAGB recognizes the commissioner's independent responsibilities.

NAGB Policies

Among the NAGB policies pertinent to the technical quality issue, we found that the "Policy on Policies" (December 1989) states that NAGB policies that address its legislated responsibilities will state the end to be achieved but not the means of policy implementation: the means will be left to the implementor. However, this restriction was not observed in the levels example: NAGB's levels policy specified both ends and means. In commenting on this observation, NAGB explained that the levels policy covers both ends and means because the identification of achievement goals is a function specifically assigned to NAGB by statute, and NAGB itself was the implementor.

The functions assigned by statute, however, are the very functions to which this policy is directed. In light of NAGB's comment, it seems

important to clarify whether NAGB is responsible only for policy with respect to these functions (with the expectation that implementation will be delegated to NCES or to a qualified contractor) or whether it is responsible for implementation as well and can choose not to delegate.

The "Policy on Policies" also requires public or expert involvement or both prior to final NAGB action and permits (but does not require) NCES, the NAEP contractor, and NAEP administrators to make suggestions concerning policy alternatives. In the levels case, we found that NAGB offered such limited information so little in advance of public hearings that technical experts found it difficult to offer useful comments. In our view, mere permission for NCES and NAEP experts to suggest alternatives is a weak use of a major technical resource.

NAGB's "Policy on Reporting and Dissemination" (November 1990) requires all NAEP reports to follow NCES review and clearance procedures and to be free from interpretations that are not supported by data; it also specifies the use of the levels framework as the primary means of reporting from 1992 on.[7] This policy also gives NAGB the power to issue companion reports that may be more speculative and interpretive than NAEP reports but must be clearly distinguished from them. NAGB has not established quality standards for its own reports.

We find that these policies and associated procedures offer only weak protection against technically unsound direction and reporting, especially since NAGB is apparently not compelled to delegate policy implementation to knowledgeable experts.

Conclusion

We conclude that NAGB's knowledge resources and procedures do not provide reasonable assurance that work done at its direction will be technically sound. NAGB's governance structure, membership, and procedures offer little protection against technically unsound policy direction for NAEP and do little to encourage strong technical input throughout policy formation. NAGB is capable of using the technical resources that are available to it but may fail to see the need or may choose not to do so.

NAGB is independent and works at its own direction. If it mandates a technically uninformed course of action with respect to a function for

[7]NAGB officials assured us that the levels will be used only if they pass NCES review; however, the policy does not state this explicitly.

which NCES is responsible, the only recourse is for the commissioner to refuse to carry out NAGB's instructions—in the case of data reporting, to follow NAGB's instructions knowing that they are likely to be overturned as a result of the prepublication review. This arrangement is wasteful in that it allows errors to be detected only after time and resources have been expended on flawed work, and it puts the commissioner in a difficult position.

Considering all the evidence, we conclude that the risk remains high that NAGB may fail to recognize—even after advice from technically knowledgeable experts—that a policy issue has critically important technical implications and, thus, may give unsound technical direction to NAEP.

Recommendations

We recommend a number of steps that NAGB should take to ensure that technical aspects of proposed policies receive early and expert attention and that the technical quality of all publications is maintained. These steps can be taken within the existing structure and do not require any change in legislation. We also recommend that the Congress review the division of functions between NAGB and NCES, with a view to aligning those functions more closely with organizational strengths and capabilities.

Recommendations to NAGB

To ensure that it does not formulate and adopt technically unsound policies or approve technically flawed results, we recommend that NAGB

1. obtain NCES review of the technical strengths and weaknesses of proposed policies that implement NAGB's statutory responsibilities, prior to final decision on such policies;[8]

2. analyze the probable effect of proposed policies (such as the achievement level policy) on NAEP's ability to present achievement fairly and accurately and to support trend reporting that is both valid and reliable;

3. pilot test and thoroughly evaluate any new design or analysis procedure before it is fully implemented and results are reported; and

[8]When a policy has implications for test design, technical experts should be involved as early as possible in the policy planning. Our recommendation is intended not to preclude NAGB from commissioning technical reviews from independent experts but simply to ensure that the experts who know NAEP are fully consulted.

4. adopt standards of technical quality (to be applied internally) for publications issued under its own authority and also secure competent external technical review of such publications prior to authorizing their release.

We recommend that the Chairman review actions NAGB has taken with respect to its statutory responsibilities in the past 2 years, identify those whose technical consequences have not been sufficiently examined, and secure technical review as necessary to ensure that these actions will not generate unanticipated technical difficulties in the future.

We also recommend that the Chairman of NAGB review each proposed policy to ensure that NAGB prescribes policy ends, not technical details of implementation.

With respect to NAGB membership, we recommend that NAGB nominate for the two testing and measurement positions only persons with relevant professional qualifications who are trained and experienced in the design and analysis of large-scale educational tests. To further add technical expertise within its currently mandated membership structure, NAGB should also ensure that two or more of its elected officials, educators, and representatives of the general public have significant technical knowledge and experience.

Recommendations to the Congress

We recommend that the Congress clarify the division of responsibilities between NAGB and NCES, with a view toward concentrating NAGB's efforts on the functions for which its broad representation is an asset and toward distinguishing functions NAGB itself is to implement from matters on which it is to give policy direction or advice to the commissioner. While NAGB as it is currently constituted can appropriately advise the commissioner from a constituency perspective regarding functions that are technical (such as the method and design of the assessment), it does not have the technical resources to carry out these functions and should be relieved of this apparent responsibility. When the Congress has more clearly determined what NAGB's functions should be, it should review NAGB's membership and determine the number of technically trained members needed.

Comments From the U.S. Department of Education

Note: GAO comments supplementing those in the report text appear at the end of this appendix.

UNITED STATES DEPARTMENT OF EDUCATION

OFFICE OF THE ASSISTANT SECRETARY
OFFICE OF EDUCATIONAL RESEARCH AND IMPROVEMENT

RECEIVED
MAR 26 1993
GAO/PEMD

MAR 25 1992

Ms. Eleanor Chelimsky
Assistant Comptroller General
General Accounting Office
Washington, D.C. 20548

Dear Ms. Chelimsky:

The Secretary has asked that we respond to the draft report titled, *Educational Achievement Standards: NAGB's Approach Yields Misleading Results*, which was transmitted on February 22, 1993. Your report is timely because the current law expires this year and the Administration and Congress will be considering appropriate legislative changes for the National Assessment of Educational Progress (NAEP). Many of the findings, comments and recommendations are directed to actions of the National Assessment Governing Board (NAGB). The Board is responding separately to your invitation for comments and is including its views on, among other things, changes in its achievement level setting process for 1992 that were not detailed in your report. These changes are also a significant part of the overall effort that the Board has been conducting to fulfill a provision of the law that directs them to identify "appropriate achievement goals" for each age and grade tested in NAEP. Our comments are addressed to those issues related to the Department of Education and its National Center for Education Statistics (NCES), as well as to more general issues about standard setting and statistics.

See comment 1.

While the General Accounting Office (GAO) report deals primarily with technical aspects of NAGB's actions to set achievement levels, in fact, the concept of performance standards involves much more than that. Any attempt to establish performance standards raises questions of substance: what it is that we want American students to know and be able to do, and how well we expect them to do it. Performance standards also raise questions of public policy: whether our national assessment should lead, or should follow, student learning progress, and how we decide, as a nation, what the standards should be. The Governing Board is attempting to set performance levels to challenge American students. The National Education Goals and the Administration's legislative proposal, Goals 2000: Educate America Act, support that position. On the other hand, the Governing Board also supports

See comment 2.

Page 2-Eleanor Chelimsky

gathering high quality data on trends in student performance. The task of balancing these two purposes is challenging, but necessary. The GAO report, however, appears only to support the limited trend monitoring role for the National Assessment. We believe that each of these roles, properly executed, can serve a constructive purpose in informing the public.

The National Center for Education Statistics is supporting several studies about standard setting for NAEP, and specifically about the achievement levels adopted by NAGB, through the National Academy of Education. When completed, these studies will help to inform the national debate about standards-based education. But the policy and substantive issues, not just the technical ones, will be critical in deliberations of Congress as it deals with the Administration's plan for Goals 2000.

1992 DATA RELEASE

The National Assessment of Educational Progress (NAEP) is a complex project conducted by NCES with policy guidance from NAGB. Among other things, NCES has the responsibility to ensure that: (1) all reports and releases of NAEP data meet accepted standards of technical soundness, and (2) the data are released in a timely manner. As you are aware, the Center has awarded a grant to the National Academy of Education to evaluate the Trial State Assessment, including achievement levels. It has also contracted with University of California at Los Angeles (UCLA) to conduct studies of many features of NAEP, including achievement levels. These studies and our decisions about appropriate follow up action are ongoing. At the same time, States have invested significant resources in NAEP and many are depending on timely availability of data for critical policy decisions. In the meantime, however, NCES has tried to balance its responsibilities for sound and timely reporting of the 1992 NAEP data with NAGB's policy directive that achievement levels be the first and primary method of reporting the 1992 results.

NCES and its contractor, Educational Testing Service (ETS), have prepared a report of the 1992 mathematics assessment in the Nation and the States using the achievement levels as one of the primary means of reporting the data. But to ensure high technical standards appropriate for a Federal statistical agency, the report also includes detailed information about means and distributions of student performance and also about anchor points—descriptions of what students know and can do at various points on the NAEP scale—similar to reports issued since 1985.

Page 3-Eleanor Chelimsky

The data will be released at a press conference on April 8, 1993, in a report titled, *1992 Mathematics Report Card for the Nation and the States*. The report will contain appropriate caveats about the achievement levels based on a number of studies, some of which are still in process. NCES will also release a report titled, *Interpreting NAEP Scales*, which will contain a full discussion of different methods of reporting and interpreting NAEP data.

Since January, NCES has taken a number of steps to invite advice about what are appropriate inferences that can be made from the achievement levels. NCES has taken the issue to NAEP's Design and Analysis Committee, to the committee that advises NAGB's achievement level process, and to a March 3 meeting with many experts and interested parties to review studies on deriving appropriate inferences about student performance with respect to the achievement levels conducted by NAGB's contractor and the Center's Technical Review Panel. At this point, it is still not clear what inferences can confidently and accurately be drawn about student performance at the achievement levels. Furthermore, it appears that the development of the descriptions and the levels themselves are still in process and may be different in 1994 from what they were in 1992.

NCES will continue to seek technical advice about the achievement levels and, in subsequent releases of the 1992 reading and writing assessment data, may issue separate "research and development" reports with the achievement levels as the means of reporting and other data reports without achievement levels.

RESPONSE TO RECOMMENDATIONS

Now on page 38.

Recommendation: (p. 2-38) In light of the many problems we found with NAGB approach, we recommend that NAGB withdraw its direction to NCES that the 1992 NAEP results be published primarily in terms of levels. The conventional reporting format should be retained until an alternative has been shown to be sound. Second, we recommend that NAGB and NCES develop a joint plan and schedule for review of NAGB's achievement levels approach, taking into account evaluations that are currently under way and providing for additional activities as needed.

We are continuing external evaluations of achievement levels. NCES views development of alternative approaches as an iterative process in which there should be an opportunity for the public to know what is at issue and participate in the conversation on an informed basis. In the meantime, we are continuing our conversations with NAGB about appropriate reporting in future NAEP reports and Center reports continue to provide information about means and distributions of student performance and anchor points.

Page 4-Eleanor Chelimsky

Now on page 51.

Recommendation: (pp. 3-21, 22) In view of the conceptual and technical flaws inherent in NAGB's achievement levels approach (see chapter 2) and of the many questions that need to be resolved before an alternative standard-setting method can be selected, we recommend that NAGB withdraw its policy of applying the 1990 achievement levels approach to future NAEP tests and join with NCES in exploring alternatives for setting both content-based and overall performance standards with respect to NAEP. This inquiry should examine issues of purpose, technical feasibility, cost, fairness, credibility and usefulness.

We agree that a collaborative effort to explore alternative ways of setting standards would be a constructive activity. It would also be timely since the national curriculum standards projects are, in various ways, exploring the same question. Eventually, this may necessitate design changes in the administration and scaling of NAEP. The approach used in the future needs to take into account both the content and performance information needs implied by standards such as those developed by the National Council for Teachers of Mathematics and those to come in other subject areas, the information needs of the National Education Goals Panel, the need to measure trends over time, and the need to monitor the performance of the Nation and the States in comparison with National Education Goals. The approach used in the future should be beyond technical reproach, and provide a way to link standards with test-based data about what students "can do" at each achievement level.

Now on page 51.

Recommendation: (p. 3-22) We recommend that the Congress specify what it intends in directing NAGB to identify appropriate achievement goals: whether it envisions the establishment of overall performance standards, the establishment of content-based performance standards, or simply better alignment of test coverage with content mastery standards. Given that legislation to establish a mechanism for adopting national content standards is currently under consideration, the Congress may also wish to express specific guidance with respect to activities to align NAEP to content standards before these mechanisms are in place.

In general, we believe that it would be appropriate for Congress to provide more specific guidance as to what it means by the "appropriate achievement goals" phrase in the law, and we will consider this issue in drafting our reauthorization proposal. In that regard, at least one possible optional definition has been omitted from this recommendation: that achievement goals might be the setting of targets for the proportion of fourth, eighth, or twelfth grade youth who should demonstrate mastery of various areas of knowledge or skills.

Page 5-Eleanor Chelimsky

Now on page 64.

Recommendation: (p. 4-23) We recommend a number of steps that NAGB should take to ensure that technical aspects of proposed policies receive early and expert attention and that the technical quality of all publications is maintained. These steps can be taken within the existing structure, and do not require any change in legislation. We also recommend that Congress review the division of functions between NAGB and NCES, with a view to aligning those functions more closely with organizational strengths and capabilities.

We concur that st· ıtory.clarification of the important roles that NCES and NAGB have to play in the NAEP project could serve a constructive purpose, and we will also consider this as we develop our reauthorization proposal. NAGB is well suited to provide broad policy advice by representing the many constituents served by the NAEP project. NCES is well suited to provide the operational and technical expertise needed to conduct a complicated survey like NAEP. Both functions are needed in order to ensure that the assessment data are technically valid and reliable and, at the same time, policy relevant and worth the expenditure of considerable public funds.

OTHER COMMENTS ON THE REPORT

Now on page 22.

As stated on p. 2-3, the NAEP test has historically been designed to *"describe the range of performance... and to measure the average accurately."* However, NAGB sought to use the NAEP scale for a different purpose: to measure student performance in terms of standards of what students at three levels of achievement... should know and be able to do. This different purpose was not ideally served by the current NAEP framework and item pool. In the future, NCES and NAGB can explore design and scaling modifications that might be needed to support achievement levels more fully.

See comment 3.
Deleted.

We disagree with your observation on p. 2-6 that measuring content would require *"measuring achievement in terms of a separate scale for each level."* There is nothing in the NAGB procedure that would require development of a scale for each level. Although there are enough items to develop six subscales (algebra, geometry, estimation, etc.), there are not enough items to develop nine achievement level scales. NAGB did not develop achievement levels for each of the six subscales because this would have resulted in 18 standards (basic, proficient and advanced for each of the six subscales) that the current item pool supply could not support.

See comment 4.
Now on page 27.

As stated on p. 2-7, it is true that there are many different patterns of answers that achieve a given percent correct. In fact, NCES and the Educational Testing Service (ETS) originally recommended that the pattern of rater-responses (not the percent correct) be used in setting the standards. This would have meant that the standards

Page 6-Eleanor Chelimsky

would have been set in the same way the test is actually scored for students. This idea was rejected by the advisors to NAGB on the grounds that the percent correct approach was (a) more consistent with the traditional Angoff procedure, (b) easier to explain, and (c) yielded the same results (on the average) as the more complicated pattern of responses. NCES believes that use of the pattern of responses would have been a better approach, but it probably would not have substantially changed the standards that were selected.

Now on pages 23-24.

The report's observations on pp. 2-7 and 2-8 that the *"NAEP tests were not designed with occupational skills or advanced college course prerequisites in mind...predictive standards cannot be based on judgment alone; they must be backed by factual information"* are true. Such issues in external validity will be considered by the National Academy of Education as part of its four external validity studies. Of course, NAEP is not meant to be predictive.

Now on page 25.

We agree with the report's observation on p. 2-12 that one reason the judges may have set such high standards is that they did not have the disciplining experience of comparing their personal estimates of what students at a given level will do with what students like those at that level actually did. Our understanding is that at the time the standards were first set, the judges did not know how well the students were doing on each item within the range of each standard. Also, the judges did not know how many students met or exceeded each standard. Had they used this type of "reality" feedback they may have set a different standard. A related issue (on p. 2-13) is that the judges were not given information about the probability of guessing the correct answer for each item. Again, such information may have affected the judges ratings.

Now on page 26.

See comment 5.
Now on page 31.

Now on page 32.

We are puzzled by the finding on p. 2-18 that basic and proficient level students did better than they should have according to NAGB's judgments on the items relevant to that level. By contrast you found (p. 2-19) that students at the advanced level did less well than they should have according to NAGB's judges. All of the evidence NCES has seen to date indicates that what students "can do" at the various levels is less than what NAGB's judges say they "should do" at each level. This led NCES to ask the NAEP Technical Review Panel to conduct several studies on this topic and to expand discussion by holding meetings on these issues.

Now on page 30.

We are not sure that your conclusion on p. 2-24 is correct when you state that the performance on advanced achievement levels "is extreme even by world standards." We cannot tell from your report what equating methodology you used. Initial results from analyses conducted by the NAE and ETS indicate that the high achieving countries on the International Assessment of Educational Progress (IAEP) did quite

Page 7-Eleanor Chelimsky

well at the advanced level. If this is true, then this implies that the advanced level is not extreme relative to the highest performing countries in the IAEP study.

Now on page 34.

We find the statement on p. 2-29 that NAGB's interpretations *"confound what panelists thought students should do with what they actually could do"* consistent with the NCES experience during adjudication of the 1992 mathematics reports. Various reviewers wanted to substitute "can do" for "should do" in the report. This led NCES to seek evidence of the extent to which the achievement level descriptions might interchangeably be used to describe what students can do.

Now on page 36.

The external validity evidence on 1992 achievement levels mentioned on p. 2-33 is under study by the National Academy of Education. Four studies on this topic (along with six studies on internal validity) are being conducted. The results will not be available until Summer 1993, however.

See comment 6. Deleted.

There is a misstatement on p. 2-35. Although NAGB's policy requires that NAEP reports meet NCES technical quality standards, the law does not specifically contain such a requirement. Such a requirement would clarify the implied NCES role and would result in an explicit sanction for NCES technical input at the design, implementation, analysis, and reporting stages of NAGB's activities.

Now on page 40.

We would find it helpful if your discussion beginning on p.3-1 about potential forms for student performance standards were expanded to include more options. For example, standards could be written into the scoring rubrics for performance assessments (NAEP has done something like this with past writing assessments). Alternatively, standards can be grounded in external criteria (such as performance on the Advanced Placement Test, ACT, or SAT) which serve as a benchmark for NAEP. A wide array of approaches to standard setting should be reviewed before decisions are made about future procedures to set standards for the NAEP test. We agree with your observations on p. 3-13 that the NAE will "provide reliable information about the feasibility and usefulness of alternative standard-setting methods."

Now on page 47.

Now on page 48.

As noted on p. 3-16, setting standards on an existing pool of items may lead to biased standards if the pool is not sufficient. In the case of the 1990 and 1992 NAEP mathematics assessments, the pool of items was representative of the content domain to the extent that NAGB's item and test specifications are explicit. If the item and test specifications were more explicit, then the item pool could more fully represent the content of the frameworks.

Now on page 49.

Now on page 50.

Page 8-Eleanor Chelimsky

We agree with p. 3-18 that NAEP cannot serve the dual purpose of providing content standards as well as overall performance standards. In fact, we feel that one of the fundamental difficulties with the approach taken by NAGB was that the achievement levels were developed with the goal of overall performance standards, whereas the written descriptions had the goal of describing content standards.

It is true, as stated on p. 3-20, that the NAEP samples could be drawn in such a way that different types of standards could be applied to different nationally representative samples. Some samples could serve the purpose of monitoring trends, others could provide multiple ways of assessing standards, and other samples could be used for a host of experimental and innovative approaches to assessing students. Such design changes will be considered in the future as NAGB and NCES work to find better ways to implement standards-based reporting.

We appreciate the opportunity to review the draft report and we hope that our comments will be helpful to you in producing the final document. Please do not hesitate to contact us if you have any questions on our comments or for more information.

Sincerely,

Emerson J. Elliott
Acting Assistant Secretary

The following are GAO's comments on the U.S. Department of Education's March 25, 1993, letter.

GAO Comments

1. The department observes that performance standards involve questions of what students should know as well as of how well they should do on the test and that our report deals only with the latter. We agree that the two issues are bound together. Ideally, deciding what students should know (and at what level of difficulty) should be the first step in a standard-setting process and should guide the design of the test. (See table 3.1.) Proceeding in this sequence helps ensure that the test and the manner of scoring are appropriate to the standards and that test scores can be interpreted in terms of the standards. In the case we studied, test content and scoring practices were in place before the standards were framed. The scores and the standards represented somewhat different concepts of performance, which led to problems in interpretation. We have added to chapter 3 to clarify that test content is important to the establishment and interpretation of overall performance standards.

2. The department comments that we appear to support only a monitoring role for NAEP. We have amended the text of chapter 3 to make clear that NAEP tests can be used to set standards of overall performance on material that all students are expected to know.

3. Our observation about subscales was not intended to refer to subscales by content area. Since NAGB's approach required the use of the NAEP scale, the point we were making was somewhat moot and has been dropped.

4. We are interested in the department's observation that using a pattern approach rather than the percent correct approach to find NAEP scores matched to the item judgments would not have made much difference in the scores selected. The pattern approach should correct for one possible source of error (that is, for differences in item weighting between the item judgment procedure and NAEP scaling). It does not correct for overestimates of basic-level student performance on very difficult items. Our analysis assumes both corrections.

5. The department remarks that our analysis of student performance on items of varied difficulty produced puzzling results. As we note in chapter 1, our analysis represents a new way of looking at student performance on NAEP. We chose this method because it seemed related to the concept of performance expressed in NAGB's definitions. We are pleased that studies

of student performance at various NAEP scores are continuing, and we look forward to learning of the results.

6. The department notes that the statute that authorizes NCES and NAEP does not specifically require that NAEP reports meet NCES technical quality standards. Upon rereading the statute, we agree with this observation. (The statute provides for the establishment of such standards and requires that NAEP reporting be fair, accurate, reliable, and valid but does not expressly state that NAEP reports must meet NCES standards.) We have deleted reference to the statute from our text.

Comments From NAGB

Note: GAO comments supplementing those in the report text appear at the end of this appendix.

THE NATION'S REPORT CARD

National Assessment Governing Board

National Assessment of Educational Progress

March 23, 1993

Eleanor Chelimsky
Assistant Comptroller General
Program Evaluation and Methodology Division
General Accounting Office
Washington, DC 20548

Dear Mrs. Chelimsky:

Enclosed herewith is the response from the National Assessment Governing Board to the draft GAO report concerning the setting of achievement levels by the Board, and the Board's use of technical expertise.

This response was approved for transmittal to the GAO by the Board's Executive Committee on March 19, 1993. Attached also, is a short paper from Ronald Hambleton, the Board's chief consultant for its 1990 effort; a paper from American College Testing, the Board's lead contractor for the 1992 levels-setting; and a paper from Gregory Cizek, an expert in standard setting, who was uninvolved in either the 1990 or 1992 achievement levels activities.

Please do not hesitate to contact me if you have questions regarding our response.

Sincerely,

Mark D. Musick
Chairman

RECEIVED
MAR 23 1993
GAO PEMD

Enclosure

800 North Capitol Street, N.W.
Suite 825
Mailstop 7583
Washington, D.C. 20002-4233
(202) 357-6938

INTRODUCTION AND HIGHLIGHTS

The General Accounting Office's draft report on achievement levels for the National Assessment of Educational Progress is based on the same misunderstandings that appeared more than a year ago in the agency's interim report. It reflects the same fundamental disagreements about the value and nature of standards for educational performance.

In summary, the National Assessment Governing Board makes these main points:

See comment 1.

- National Assessment results should be reported primarily in terms of challenging standards that help the nation determine "how good is good enough." The conventional practice of simply comparing one group of students to another is no longer adequate. GAO makes no compelling argument for returning solely to the older methods of reporting by means, percentiles, and "benchmarks."

See comment 2.

- The Board and numerous other groups believe that achievements levels can properly be used to report results on the National Assessment. We reject the argument that trying to set standards on NAEP is "conceptually flawed." We reject GAO's recommendation that the 1992 achievement levels be withdrawn.

See comment 3.

- The GAO report is unbalanced and misleading. Many of its assertions are undocumented; much of its analysis is flawed.

See comment 4.

- The GAO report is out-of-date. It focuses on the achievement levels for 1990--indeed, mostly on the first phase of the process for setting them which did not form the basis for the levels actually adopted. It gives relatively little attention to the standard-setting process for 1992 and fails to recognize the improvements made.

The process for setting the 1992 achievement levels was conducted under a $1.5 million contract by American College Testing (ACT), which has extensive experience in standard-setting in many fields. ACT consulted regularly with a panel of leading experts in measurement and standard-setting who believe strongly in the feasibility of setting standards on NAEP and in the soundness of the process used to advise the Board on what the levels should be.

See comment 5

The movement from norms, based on test averages, to standards, based on informed judgment of what students ought to know and do, is occurring not only on NAEP but in many parts of American education. It stems from dissatisfaction with "national norms," which by definition place half of all students below an average score that may be woefully inadequate. The movement to standards also reflects the conviction that setting clear markers of what students should learn makes any test far more useful and meaningful to parents, schools, and the public.

See comment 6.

Yet, the authors of the GAO report seem cool to this central idea. They frame the issue as "statistical quality," not policy judgment. They suggest alternatives that would not really yield standards at all, just norm-referenced descriptions of performance. For example, the Board rejects the kind of "benchmark" example suggested by GAO in which acceptable performance is arbitrarily set at the 30th percentile of student achievement.

The report seems premised on two major misinterpretations. First, it fails to recognize the extent to which setting test standards involves policy judgment rather than a technical process to find an "accurate" score. Second, in contrast to what the report asserts, standards often are set on tests quite similar to NAEP using the same system of collecting judgments--the Angoff procedure. Far from being "novel," the procedure is widespread.

In arriving at the standards, most of the experts on whose judgments NAGB relied were classroom teachers, bringing first-hand experience from many parts of the nation. The standards adopted contain reasonable descriptions of what students should learn. They are meant to denote overall levels of proficiency, well-suited for placement on the NAEP scale, not checklists of specific skills.

The GAO report relies on outmoded models of psychometric evaluation. In particular, it conceives of validity as an all-or-nothing proposition when it properly is a matter of degree, based on the weight of the evidence and the uses made of results.

NAGB believes that using standards on NAEP is a developing process. It has adopted preliminary descriptions of the levels as part of the frameworks for 1994 NAEP exams, and is certain there will be other changes over the years to make achievement standards a primary factor in creating NAEP assessments as well as in reporting them. It believes strongly, though, that any improvements that may occur in the future do not detract from the overall soundness and utility of the 1992 NAEP achievement levels being developed by ACT.

The Governing Board agrees with GAO about the importance of securing technical advice, and has done so regularly in regard to achievement levels, as well as in its other work. However, because of the wide impact of NAEP, the assessment should be guided by an independent, widely-representative policy-making board--not a closed circle of federal officials and technicians.

Appended are comments by ACT; Ronald Hambleton, of the University of Massachusetts; and Gregory Cizek, of the University of Toledo.

FROM NORMS TO STANDARDS: A BASIC POLICY ISSUE

There is an issue of basic policy underlying both GAO's report and NAGB's comments:

Should the National Assessment present results in terms of performance standards rather than simply using scale scores and proficiency levels, based on the distribution of test results? Should NAEP include judgmental standards of what achievement ought to be?

The Board believes standards are essential, and it is responsible for setting them under NAEP's 1988 authorizing legislation. The law calls for "appropriate achievement goals for each grade...and subject area to be tested under the National Assessment," and these are clearly meant to differ from NAEP's previous descriptions of actual performance.

By giving an informed, deliberate judgment about "how good is good enough," the achievement levels make the National Assessment far more useful and meaningful to policy-makers and the public. They also help NAEP play its crucial role in tracking progress toward the national education goals. It is no longer enough, the Board believes, for NAEP simply to report who is above average and who is below without giving a sense of what should be expected.

A similar interest in reporting by standards rather than by norms is at the heart of reforms in many state testing programs and of the New Standards Project. Yet, the authors of the GAO report seem cool to the idea. They frame the issue as "statistical quality" not policy judgment and repeatedly counsel "reasonableness" and "realism," apparently code words for not expecting too much.

See comment 7.
See comment 8.

Most of the report's suggested alternatives for standard-setting would simply describe the performance of various groups, not deal with the substance of what students should know and be able to do to move beyond the status quo. Although the report ostensibly takes a technical stance, its authors' policy views clearly color their descriptions, interpretations, and conclusions.

TWO MAJOR MISINTERPRETATIONS

The report seems premised on two major misinterpretations of NAGB's standard-setting effort. First, it fails to recognize the extent to which setting standards on any test involves policy decisions and judgments rather than a tightly-defined technical process to find one "accurate" score. Second, in contrast to what the report asserts, standards often are set on tests quite similar to NAEP through the same system of collecting judgments--the Angoff procedure--that NAGB used.

See comment 9.
See comment 10.

The standards on NAEP, as on similar exams, denote a general, overall level of proficiency; they are not intended to be a specific check-list of skills. The standards are valuable for interpreting the significance of NAEP results--far more valuable than any skill check-list could ever be.

See comment 11.

Repeatedly, the report complains that the achievement levels are not "accurate" and do not "measure" what they are supposed to. In fact, the measuring in NAEP is done by test items and the scale used in reporting results. The achievement levels, like any standards, are a series of judgments about how to interpret these results. It is reasonable to debate whether they are too high or too low, but to say they are "not accurate," as if some "true" levels exist, is absurd.

See comment 12.

As Richard Jaeger writes, "All standard-setting is judgmental. No amount of data collection, data analysis, and model building can replace the ultimate judgmental act of deciding which levels of performance are meritorious or acceptable and which are unacceptable or inadequate." (Jaeger, R.M., "Measurement Consequences of Selected Standard-Setting Models," 1979.)

See comment 13.

The central issues that should be considered are whether the judgments themselves are informed and defensible and whether they are arrived at through a careful process. For NAEP the judgments ultimately are made by the Governing Board, but they are informed by a wide, public consultative process, as well as by the careful judgments of broad-based panels of experts, most of whom are classroom teachers.

The Board believes that the achievement levels in mathematics contain reasonable descriptions of what students should learn in the different grades. GAO does not challenge this view, and, as the report even notes, the levels for 1990 received strong support from educators and policy-makers who considered them, as shown in surveys conducted for NAGB.

See comment 14.

Far from being "novel," as the report claims, the Angoff procedure to apply expert judgments to the NAEP exam is the most widely-used standard-setting process in the nation. It has been implemented on hundreds of tests over two decades. The first to suggest using it on NAEP was Albert Beaton, one of the leading theoreticians in the development of NAEP for Educational Testing Service. The specific design for 1990 came from national experts in measurement. As the process developed, it also took into account recommendations by a panel of state testing directors.

The design for 1992 was prepared by ACT, which has used variants of this system on dozens of examinations. Again, there was wide input and review by national experts, including staff of the National Center for Education Statistics. To suggest, as GAO does, that NAGB, a largely non-technical, policy group, took it upon itself to develop a standard-setting process is simply not true. Yes, the Angoff procedure was "modified;" it nearly always is adapted to specific situations.

VALIDITY

In discussing validity, the GAO report is confused, using an outmoded "all-or-nothing" approach. For example, the report says "valid measurement depends, ultimately, on having a measuring instrument and a method suited to the purpose." Yet, the most influential recent work on the subject, by Samuel Messick, of ETS, explains that validity concerns "not the test or observation device as such but the inferences derived from test scores or other indicators."

See comment 15.

As Messick defines it, validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment." He continues, "Validity is a matter of degree, not all or none....[It] is an evolving property and validation is a continuing process." (Messick, S. "Validity" in Educational Measurement [3rd. Edition], 1988) Thus, the GAO report misstates an important point when it says validity is a property of the test and not a characteristic that applies to the interpretations of test scores.

When it discusses whether NAGB followed "a valid process" in setting achievement levels, virtually all material in the report deals with the process used for 1990, and even that is limited. When it deals with valid inferences from the results, all of its evidence relates to 1990--not the levels for 1992. A number of validity studies of 1992 achievement levels are underway. Such studies should continue and more should be done in future years. The Board expects the information they provide will be useful in developing standards for future NAEP assessments.

SIGNIFICANT CHANGES FOR 1992

See comment 16.

As explained in the attached response by ACT, there were significant changes in how the Angoff procedure was carried out on the 1992 assessments, compared to 1990:

- more opportunity for public input
- increased technical support
- an elaborate procedure for selecting judgment panels which assured broad representatior nationwide
- development of operational definitions before individual test items were rated
- increased training and feedback for panelists
- increased time for panel deliberations
- a pilot study of the entire process
- a built-in reliability check by splitting the item pool between groups of judges

These seem to have met many of the specific concerns that GAO has raised. Yet, the report only briefly examines the changes. Instead, it simply asserts that no matter what improvements may be made in procedure, NAGB's approach is "inherently unsound" because the nature of the National Assessment precludes it.

The reason for this "fundamental flaw" is not clearly explained in the report. It seems to be that performance standards which refer to content mastery can only be properly set on a specially-designed criterion-referenced test. An example might be the "mastery exams" used in many grade-school classrooms to determine if a particular type of addition problem has been learned, although GAO does not spell that out.

It is true, of course, that there can be no skill-by-skill confirmation of what students know unless they take a test designed to give skill-by-skill results. NAEP can't do that--nor can most widely-administered exams. But that's not what NAGB's achievement standards are trying to do.

The central language of the standards, used in the definition of proficient, is taken directly from the National Education Goals: that students should have "demonstrated competency in challenging subject matter." This is the "solid academic performance" which the proficient level is intended to represent. Clearly, the achievement standards, like the Goals, refer to a general degree of attainment. That is what NAEP is meant to show, as GAO explains.

The text of the subject-matter descriptions which are part of the achievement levels is written in general terms to describe the sorts of skills and knowledge that students at each level should have. It is derived from the test framework and is useful, as GAO recommends, in giving judges a common basis for rating test items and in explaining to the public the level of competence expected.

See comment 17.

Contrary to what the report asserts, the achievement levels are meant to be a general performance standard on the NAEP scale. NAGB is not trying simultaneously to make detailed specifications of content mastery. For GAO to say so is based on a strained misreading of Board policy, not on the achievement levels as they have actually been developed. In essence, the report creates a straw man that it can easily knock down; thus, much of its lengthy discussion is irrelevant.

See comment 18.

What is relevant is that most standard-setting in American schools and professions is done on tests, like the National Assessment, that contain a wide range of items relevant to the standards being set. In some cases the standards themselves may be part of the test design, and NAEP may evolve in that direction. But this type of test development is not a precondition for meaningful standards. As a matter of fact, what the Board has done in setting overall performance standards is analogous to the five levels of performance on Advanced Placement tests and to what millions of teachers do every day in grading examinations.

Again, although one may criticize how NAGB went about setting standards on NAEP, there is no "fundamental flaw" in setting them. Indeed, neither ETS nor ACT--two of the nation's most prestigious testing organizations--has ever brought to the Board's attention the point that GAO makes even though both were involved in developing the methods for setting the standards and for placing them on the NAEP scale.

RELIABILITY

In its very brief treatment of reliability, the report makes no distinctions among the various types of error that may be part of a large-scale assessment. There is error associated with the assessment itself, error associated with the sample of students being assessed, error associated with a standards panel's judges, and finally, error associated with the estimates of the proportion of students that meet each standard.

Setting achievement levels has no impact on the first two sorts of error estimates. These exist with or without standards and to the same degree.

See comment 19.

Regarding the judges' estimates for 1990, no definitive estimates of error can be made because of the lack of two independent samples of judges' ratings. Therefore, whatever conclusions the report has reached can only be surmises, for which no one has data to prove or disprove claims about reliability.

Although reliability data for 1992 are not yet complete in all subjects, analysis so far indicates reliability in math is high, according to ACT's technical report. For 1992 the item pool was divided in half, permitting comparisons between panels of judges.

FROM JUDGMENTS TO THE SCALE

See comment 20.

One particular point in the GAO report is baffling. The report asserts that an "inappropriate method" was used to transform the judgment panels' recommendations into scores on the NAEP scales. Yet, for 1990 the method of going from an expected percent correct to a scale score was developed by ETS which created the scales; for 1992 it was done by ETS and ACT, both of which concurred in what to do. GAO presents no evidence that the procedure yields incorrect results.

This part of the process is technical work; it was carried out in close consultation with NCES. For GAO to dispute it with virtually no explanation or analysis is unwarranted.

See comment 21.
Now on page 31.

Also, the table GAO has prepared (on page 2-18), purporting to show a discrepancy between item-by-item judgments of the panelists and the actual 1990 results, is flawed. It relies on extrapolations rather than actual data, and is based on a mistaken premise, discussed above, that achievement levels are meant to be skill-by-skill specifications of mastery rather than general descriptions of degrees of mathematics proficiency. Since the goal was to set standards on the NAEP scale, it was appropriate for judges to rate all items contributing to the scale, not just those "at the level."

See comment 22.

See comment 23

Even so, there is considerable agreement between the judgments and results, which tends to confirm the process NAGB used, not discredit it. The small differences that do appear seem to reflect a regression to the mean--with students at the top scoring somewhat lower than expected and those near the bottom doing better.

Of course, there should be a good match between the judges' item judgments and the actual pattern of results because the judges knew the overall percent correct for each item as they made their judgments. In effect, easier content tied to easier questions is the basis for the lower achievement levels; harder content tied to harder questions is the basis for the higher achievement levels. Both the descriptions and the degree of difficulty of the questions follow the logical development of the subject being tested.

See comment 24.

Throughout the report, GAO seems to assume that descriptions of overall performance standards should make no reference to the particular kinds of performance that might be expected of those who meet them. This is mistaken. As Michael Kane, of the University of Wisconsin, has pointed out, in licensing and competency exams such references are usually made, just as they are in NAEP. The intent is not to guarantee skill-by-skill mastery but to provide what every standard needs--a clear indication of the level of competence expected.

Now on page 45.

In part of its report, GAO seems to understand this point well. It even suggests using the Angoff method (on page 3-12) to set "overall performance standards" on NAEP, provided there is "clear...specification of what students at each level should be able to do." That essentially is what ACT did in recommending the levels to NAGB for 1992. Yet, the report continues to reject them.

SCALING ON NAEP

The report raises two questions related to NAEP scaling. First, how can the standard-setting process include **all** items yet "arrive at a standard that validly reflect(s) mastery of only **some** items?" And second, doesn't it matter whether or not students get a certain percent correct by the same pattern of answers expected by the standard-setting panels for a particular level?

See comment 25.

The implication of both is that NAEP scaling procedures are faulty. But in fact, NAEP scaling works in such a way that it is **always** possible to relate a percent correct score to a NAEP scale point through a mathematical relationship called the "test characteristic curve." If GAO had evidence that NAEP scaling is flawed, these might be reasonable concerns. However, there have been numerous studies demonstrating its soundness and no published evidence to the contrary.

The report also raises a question about the kind of items--easy or hard--that students at the basic level are expected to answer. This, too, is a scaling issue, not an evaluation of the achievement levels First, NAEP does not provide scores for individual students, and second, no student takes all test items. Of course, there are virtually infinite patterns of answers that could lead to a particular score. But on average--and that is what NAEP scaling allows--a score of 255, for example, is most likely to reflect the progression of easier to harder questions used in setting the eighth grade basic level at that point.

See comment 26.

Many of these issues apply to the descriptive anchor points as well as to the judgmental achievement levels. Yet, the report praises anchor points as "straightforward" and "serviceable"--without analyzing how they are developed--while criticizing achievement levels, placed on the same scale, as "misleading."

ARE THE STANDARDS TOO HIGH?

See comment 27.

Despite the report's lengthy critique of NAGB's standard-setting method, one of its principal objections seems to be that the standards themselves are "unreasonably high.". The type of standards that should be set, the report says, must be "reasonable and appropriate for the general student population of the United States."

See comment 28.

It gives one example of what it has in the mind: the test of General Educational Development (GED) required to qualify for a high school equivalency certificate. For that exam the passing score is set at the 30th percentile of a national sample of high school seniors.

The standards on NAEP are not that kind of norm-referenced standard. Deliberately, they are not based on current performance, but, like the national education goals, are meant to be "challenging," to serve as goals of what achievement ought to be.

On the NAEP math exams of 1990 and 1992 only about 15 to 25 percent of students in the national samples reached the "proficient" standard; just 1 to 4 percent were "advanced;" and slightly less than two-thirds met the partially proficient standard called "basic." As GAO notes, in recent years about 60 percent of American high school graduates have gone on to college. Yet, more than half of them drop out without ever graduating.

Almost by definition, if a standard is challenging, relatively few students can reach it now. For example, in Kentucky's new statewide exams, only about 10 percent of students reached the proficient level in any subject. In Maryland about 20 to 25 percent made the "acceptable" level of 3 or better. In the first pilot tests of the New Standards Project--even though the reliability of scoring was less than desired--only about 25 percent of students reached the criteria for passing.

Each of these tests was scored independently of the National Assessment. Yet, each was committed to high standards. And each had quite similar results. In their top categories--similar to NAEP's "advanced"--the proportion of students ranged from 5 percent down to less than 0.5 percent.

See comment 29

By GAO's calculations, only 10 percent of 13-year-olds in Taiwan and 5 percent in China and Korea can reach the 1990 NAEP advanced level in 8th grade math. The American figure was 1 percent, indicating that ten times as many Taiwanese youngsters seem to have met the standard as Americans. Yet, if the U.S. is to be "first in the world" in math and science by the year 2000, as the National Education Goals proclaim, surely as many American students should

reach this level as those in Asia. To call the standard "extreme," as GAO asserts, is unwarranted. And, of course, that is a policy judgment which might properly be made by a policy board--not a "technical issue."

Also, it is uncertain whether GAO's calculations are correct. A full-scale study linking NAEP to the International Assessment of Educational Progress is scheduled for release later this year.

See comment 30.

Of course, one way to get the external validity GAO seems to want would be to set standards based on actual performance--perhaps tying advanced to the top 10 percent of students and proficient to the top half or third. But an approach such as that would not really yield standards at all, just norm-referenced descriptions of performance. The proportion to exceed a standard would be known in advance because the standard itself would guarantee the proportion.

In that case there would be no reason for anyone to deliberate about what students should know; the only decision to make would be what proportion should be given what label. There would still be the task of describing what students do know at the pre-selected points on the curve of results--a job surely for experts, but with no credible claim to be setting standards.

THE ROLE OF THE GOVERNING BOARD

See comment 31.

The National Assessment Governing Board receives a great deal of technical advice and certainly is not isolated from testing experts, as GAO suggests. Indeed, to set achievement levels for 1992 the Board contracted with ACT, which, in turn, was advised by a panel of experts who helped shape procedures and strongly endorsed them. Also, two members of the Governing Board itself are testing experts, appointed to four-year terms last October.

But the Board, which by law must be bipartisan, also includes governors, state legislators, school superintendents, teachers, and members of the public. And it is the Board as a whole that must sift through all the advice it receives to make policy judgments. That is the proper role of a broadly-representative, policy-making Board.

Creating such a board for NAEP was one of the principal recommendations six years ago of a major study commission, whose members included prominent education policy makers and psychometricians. The commission also called for state-by-state NAEP testing and achievement goals. It said such an expansion would be widely accepted only if the governance of NAEP "reflect(ed) an array of education, measurement, and policy perspectives."

Having an independent board set policy for NAEP is very much in the traditions of American education and democracy. In virtually all parts of the country lay school boards determine policy for state education departments and local schools, not commissioners or superintendents. Members of all these boards lack the expertise of full-time professionals. That

is precisely why boards are created--to represent the various publics involved, to give legitimacy to decisions, and to bring a much broader view to bear than the interests of particular professions.

If NAEP had remained low-impact and low-profile, it might well continue to be run by federal officials and contractors. But as the importance of the National Assessment grows, the role of its Governing Board becomes more crucial.

RESPONSE TO SPECIFIC GAO RECOMMENDATIONS

1. The Governing Board should withdraw its direction to the National Center for Education Statistics that the 1992 NAEP results be published primarily in terms of [achievement] levels. The conventional reporting format should be retained until an alternative has been shown to be sound.

See comment 32.

Response: The Governing Board does not agree that the 1992 achievement levels should be withdrawn. The GAO recommendation is unwarranted.

The report asserts that there are such "inherent flaws" in the procedure used to set the levels that no amount of improvement is sufficient. As our comments indicate, GAO has misunderstood the nature of standard-setting on NAEP and applied inappropriate criteria and methods in evaluating it. Further, the Angoff method used by NAGB is the most widely employed and evaluated standard-setting procedure in the nation. In response number three below, we cite the unanimity of expert opinion given to the Board that the Angoff method is appropriate for use on NAEP. The attachment from American College Testing (ACT), our achievement levels contractor, describes changes that have been implemented on the 1992 assessments.

The achievement levels are part of a desirable shift, underway in many parts of American education, from norm-based to standards-based testing. They will improve the usefulness of NAEP to the public and policy-makers.

The Board has no objection to the second part of the recommendation that NAEP's "conventional" reporting formats, presumably averages and anchor points, continue to be used. NAGB has never maintained that achievement levels be the only way of reporting NAEP results.

We note, however, that in supporting continuation of the anchor points GAO gives approval to an approach which it has not evaluated. The report fails to consider criticisms of this approach that have been made in educational measurement literature. Even in using the term "conventional," the GAO implies that the validity of the anchor points has been established when in fact it has not.

2. The Governing Board and NCES should develop a joint plan and schedule for review of the achievement levels. This should include a determination by the Commissioner of whether (1) the approach is so technically and conceptually flawed that its results should not be proposed for publication, or, (2) the approach is sufficiently promising that preparations for NCES pre-publication review should be designed and implemented.

See comment 33.

Response: NCES has been and will continue to be closely involved in the process of setting achievement levels. Reports containing the 1992 achievement levels have gone through the NCES pre-publication review. This review has dealt with reporting issues, including the question of what inferences can properly be made about what students can do based on the achievement levels.

However, setting standards is primarily a matter of policy judgment, not of statistical quality. Determinations in this area should be made by NAEP's policy board, not by a statistical agency.

3. The Governing Board should withdraw its approach of applying achievement levels to future NAEP tests and join with NCES in exploring alternatives for setting both content-based and overall performance standards with respect to NAEP.

Response: The Governing Board, together with NCES, has already explored alternatives for setting achievement levels and will continue to do so, but it believes the Angoff method has worked well.

See comment 34.

From the initial planning to the present, both NCES and Educational Testing Service, the NAEP contractor, have been involved in the achievement levels process. In fact, the late William Angoff, then a senior research scientist at ETS, was consulted and indicated that this method would be suitable.

At a meeting in December 1991, jointly sponsored by NAGB and NCES, other approaches for setting achievement levels were explored. The view of the group--which included psychometricians from ETS and ACT and other testing experts--was unanimous: While other approaches exist, the Angoff method is the most widely used and thoroughly evaluated; no other approach is better overall for setting achievement levels for NAEP.

See comment 35.

The Board will continue to be open to consideration of other standard-setting approaches. Unfortunately, those mentioned in the report appear naive and unsupported by research evidence. However, NAGB believes that over time achievement standards should become a primary factor in preparing NAEP as well as in reporting it. The Board will not rule out proposals for other approaches in future competitions to select an achievement levels contractor.

4. (a) Congress should specify what it intends in directing the Governing Board to identify appropriate achievement goals: whether it envisions the establishment of overall performance standards, the establishment of content-based performance standards, or simply better alignment of test coverage with content mastery standards.

(b) Congress may also wish to express specific guidance with respect to aligning NAEP with any national content standards as they come into existence, given that it is considering legislation to establish a mechanism for adopting national content standards.

See comment 36.

Response: As Congress considers part (a) of this recommendation, the Governing Board submits that the distinction made by GAO between "overall" and "content-based" performance standards is drawn more sharply than warranted in practice. The report assumes, incorrectly, that overall performance standards make no reference to the kinds of performance that might be expected of those who meet the standard. In the standard-setting field, overall performance (i.e. a "passing" score on a test) is almost always associated with a clear indication of the level of competence expected. The achievement levels include particular scores on the NAEP scale, general descriptions of the content represented by that level, and illustrative test items. "Overall" and "content-based" standards are not mutually exclusive in their application, as GAO maintains. No fixed definitions should be enacted in law that would preclude flexibility and improvement.

(b) The Board believes that the content of NAEP should reflect both current and evolving instructional practice. It should neither be determined by nor ignore voluntary national conten' and performance standards as they are developed. Instead, a balance should be achieved, through the national consensus process used in developing assessment frameworks, that **appropriately** aligns the National Assessment with the standards over the course of successive administrations of a subject area assessment.

5. To ensure against technically unsound policies or technically flawed results, GAO recommends that the Board:

   (a) Obtain NCES review of policies proposed by the Governing Board prior to final decision;

   (b) Analyze the probable effect of proposed policies on NAEP's ability to present achievement fairly and accurately and to support valid, reliable trend reporting;

   (c) Pilot test and thoroughly evaluate any new design or analysis procedure before it is fully implemented and results reported; and

   (d) Adopt standards of technical quality (to be applied internally) for publications issued under its own authority and secure competent external technical review of such publications prior to authorizing their release.

6. GAO recommends that the Governing Board review actions it has taken with respect to its statutory responsibilities in the past two years, identify those whose technical consequences have not been sufficiently examined, and secure technical review as necessary to ensure these actions will not generate unanticipated technical difficulties in the future.

7. GAO recommends that NAGB review each proposed policy for conformity to its "Policy on Policies" to ensure that the Board prescribes policy ends, not technical details of implementation.

Response: Although the Governing Board does not object to the general direction of these recommendations, it rejects the implication that they are based on substantiated findings of failure on its part with respect to technical matters.

See comment 37.

The GAO report includes three studies of the Board's handling of technical issues. In two cases it commends NAGB's actions, finding that it recognized the need for technical advice and properly considered it. In only one case--the 1990 achievement levels--does GAO conclude that the Board failed to recognize that it needed technical advice. In fact, however, extensive technical advice was sought and obtained from inception to conclusion of that project. For 1992, the Board contracted with ACT, a respected standard-setting organization, to conduct the achievement levels process. Also, NCES and ETS have been closely involved throughout the achievement levels process. The assertion that NCES and ETS have been systematically ignored and uninvolved is untrue.

See comment 38.

The report is mistaken in suggesting that the Board has not properly followed its "policy on policies" with respect to achievement levels. Since setting "appropriate achievement goals" is a direct responsibility of NAGB under the statute, the Board is responsible in this case for both policy **ends** and **means**. Its "policy on policies" is intended for activities carried out by NCES and the NAEP contractor where it is appropriate that only the ends be prescribed. It is illogical to suggest that this policy should also apply to the Board in carrying out its own specified duties.

8. GAO recommends that NAGB nominate for the two testing and measurement positions only persons with relevant professional qualifications, who are trained and experienced in the design and analysis of large-scale educational tests. To further add technical expertise within its current membership structure, NAGB should also ensure that two or more of the elected officials, educators and representatives of the general public appointed to the Board have significant technical knowledge and experience.

Response: The Governing Board will continue to solicit recommendations from appropriate individuals and groups and will nominate well-qualified persons for Board vacancies as they occur. Professors Jason Millman and Michael Nettles, the members serving in the positions for testing and measurement experts, are eminently qualified. They were chosen last October, as required by law, from a list of qualified individuals nominated by the Board.

The Board seeks strong nominees in all categories of representation, and regularly solicits recommendations from more than 700 organizations and individuals. Many of its members have considerable experience with testing issues and all become well-informed. It would be unwise, however, to limit the representativeness of other categories of membership, e.g., the general public, to provide more representation in a category already having two specified positions on the Board.

9. Congress should clarify the division of responsibilities between NAGB and NCES, with a view toward concentrating NAGB's efforts on representational functions for which the Board is well designed. While NAGB as it is currently constituted can appropriately advise the Commissioner from a constituency perspective regarding functions that are technical (such as methods and design of the assessment), the Board does not have the technical resources to carry out these functions and should be relieved of this responsibility. When Congress has more clearly determined what NAGB's functions should be, it should review NAGB's membership and increase the number of technically trained members as needed.

See comment 39.

Response: As Congress considers this recommendation, the Board submits that GAO has drawn a conclusion about the capability of NAGB that is not supported by the facts. Although it criticizes NAGB in respect to achievement levels, GAO gives short shrift to the two other policy areas in which the report commends Board actions. It also ignores the consistent success of one of our major policy/technical responsibilities--conducting the national consensus process by which the framework for each assessment is developed. Certainly, planning for assessments in subjects such as reading and history provides fertile ground for controversy, yet no mention is made of the Board's positive record in these areas.

See comment 40.

Also, the report makes too stark a distinction in suggesting that technical and policy issues are discrete and easily separable. In truth, both policy and technical matters are involved in almost every issue of importance to NAEP. A sharp division between them would be unwise if NAEP is to have a governance structure with the strong checks and balances needed to maintain its independence and integrity. Because the impact of the National Assessment is wide, it is essential that NAEP have a strong, independent policy-making board, representing the wide range of interests it affects--not a weak advisory committee, as GAO suggests.

See comment 41.

The following are GAO's comments on the National Assessment Governing Board's March 23, 1993, letter.

GAO Comments

1. NAGB misstates our position as advocating "returning solely to the older methods of reporting" as opposed to reporting in terms of performance standards. In fact, we conclude in chapter 3 that overall performance standards can usefully be established for NAEP. We recommend that conventional descriptive NAEP reporting be retained for now simply because no satisfactory standards-based alternative is yet available.

2. NAGB again describes our position incorrectly. We do not argue that trying to set standards on NAEP is conceptually flawed: we find that NAGB's particular approach to doing so is conceptually flawed. As we explain in chapter 2, the conceptual flaw is that NAGB tries to do two things at once. It establishes standards of overall performance on a broad-based test but seeks to interpret the resulting scores as evidence of what students should know, not just of how well they should do on the test as a whole.

3. This general criticism summarizes specific comments on later pages. We respond to these comments in items 7 through 31.

4. NAGB's comment that our report focuses primarily on the early phase of the 1990 standard-setting is in error. Both the summary description of NAGB's approach in chapter 1 (see table 1.1) and the analysis in chapter 2 are based on the procedures actually used in setting the 1990 standards. We have added a footnote early in chapter 2 to make this clear. The report also considers the 1992 process and credits the improvements made. We have added to our discussion of the 1992 procedures to incorporate new information provided in NAGB's comments.

5. In this and the following paragraph, NAGB argues the importance of setting standards that are based on what students "ought to know and do" and on "setting clear markers of what students should learn" and remarks that we seem "cool" to this idea. We do not oppose basing the setting of standards on what students ought to know (what we call content-based performance standards). We conclude, however, that NAGB's approach is not suited to this purpose. NAGB itself seems to deny that it was its purpose to set standards of what students ought to know (see items 14 and 15): it says the achievement levels represent general degrees of attainment. This ambiguity concerning what the achievement levels represent lies at the heart of the problem with NAGB's approach.

6. The remaining material on this page summarizes points NAGB presents in more detail on subsequent pages. We respond to these points in connection with their later treatment.

7. NAGB criticizes us for using "statistical quality" as a criterion for evaluating a policy judgment. We continue to believe that technical soundness (we do not use the term "statistical quality") is pertinent. When policy judgments are based on data as NAGB's were, the quality of those data and the adequacy of the measurement procedures that underlie them are a legitimate concern. If the data are of poor quality or are based on a measurement procedure ill-suited to the purpose it is intended to serve, judgments based on them will be poorly informed and may lead to unwarranted conclusions and interpretations.

8. NAGB is incorrect in inferring that the use of words such as "realism" is evidence of our preference for not expecting much from students. We take no position respecting what should be expected of students. Rather, we take the expectations stated in NAGB's levels definitions as a given. We do use the word "realistic" with reference to the item judgments (chapter 2). Our concern is that the score selected for each level be a realistic (that is, well-informed) estimate of how students actually at that level would be likely to perform. We use the word "reasonable" (in the meaning of sensible or valid) to apply to the interpretation given to the test scores selected as standards. These usages are drawn from the literature on item judgment methods.

9. NAGB asserts that we fail to recognize that setting standards on any test involves policy decisions rather than simply a technical process that finds an "accurate" score. This misstates our position on two counts. First, we state clearly in chapter 3 that standard-setting is a matter of informed judgment. Second, we argue that such decisions should not rely solely on a technical procedure (no matter how well designed). The scores that emerge from that procedure should themselves be judged in the light of a variety of evidence to ensure that they are "accurate" in the sense that they represent the kind of performance the decision body has in mind.

10. We do not assert that the use of Angoff procedures is uncommon or that these procedures could not be applied to NAEP. We find weaknesses in NAGB's particular application of those procedures and interpretation of the resulting test scores.

11. The question of whether the NAGB standards denote a general level of proficiency or whether they also denote what students at each level know and are able to do is the central question of interpretation to which our report is addressed. We do not treat NAGB's achievement level descriptions as a "specific check-list of skills."

12. We agree that standard-setting is a matter of informed judgment. We consider NAGB's judgment insufficiently informed because (a) judgment panelists lacked necessary information and (b) NAGB did not examine whether the scores it selected could validly be interpreted as representing the achievement levels it described.

13. The question is not (as NAGB states) whether the item judgment process was defensible or the descriptions of what students should know were appropriate: it is whether those scores represent that knowledge. We found the preliminary evidence unconvincing, and NAGB has not offered evidence of its own. NAGB consulted experts with respect to the item judgment results and the knowledge paragraphs before the NAEP scores were selected, at a point when the question of interpretation could not be addressed.

14. NAGB argues here that the achievement levels approach was a sound adaptation of a common method. We do not claim that the Angoff method is novel. NAGB's use of it to apply the achievement levels definitions to NAEP was novel. (Albert Beaton recommended a quite different variant, much earlier.) NAGB asserts that it did not develop the standard-setting process. However, the record shows that NAGB did make the decisions that turned out to be critical to its approach. NAGB defined the achievement levels, determined that panelists should make prescriptive rather than realistic judgments, specified that the resulting NAEP scores should be interpreted in terms of skills that students should have, and decided to proceed without seeking to validate its interpretation. NAGB has not disputed these facts.

15. NAGB is correct in observing that thinking about validity has evolved rapidly since the professional standards we cited were written and that Messick presents the current view. We believe the substance of our draft report was consistent with that view, although we used older terminology. (For example, we asked whether NAGB's measurement method suited its purpose, which was to represent the achievement levels as NAGB defined them. Messick would ask whether there is a theoretical rationale for supposing that NAEP scores can be interpreted in terms of these

definitions.) We have revised the material in chapter 2 to make clear that our focus is on validity of inference.

16. NAGB believes that improvements in item judgment procedure for 1992 have answered our concerns. We note in our report that item judgment procedures for 1992 were improved. However, while these improvements may lead to more reliable score estimates, they will not solve the problem of interpretation. In fact, NAGB's insistence that test scores be interpreted in terms of panelists' operational definitions of what students should be able to do—without any reference at all to what students at those scores actually can do—raises new problems.

17. On this page, NAGB describes the achievement levels as general degrees of attainment and states that its intent was to set general performance standards on the NAEP scale, not detailed specifications of content mastery. This is a useful clarification of NAGB's intentions. Had NAGB presented the achievement levels in this way (that is, simply as its judgments of "how much is good enough" to be considered marginal, solid, and superior performance on the NAEP test), validity issues would not have arisen. However, NAGB sought—and by the evidence of earlier comments still seeks—to interpret the achievement levels in terms of what students should know (see item 5 above). We treat NAGB's interest in the mastery dimension of the achievement levels seriously because NAGB itself has done so.

18. NAGB draws an analogy between the achievement levels and the levels associated with advanced placement exams. As illustrated by the subject of calculus, the advanced placement levels and the process by which they are set are in fact very different from NAGB's levels and procedures. Each advanced placement test reflects a specific course outline and assesses only relatively advanced skills. The five levels or grades are (1) no recommendation, (2) possibly qualified, (3) qualified, (4) well qualified, and (5) extremely well qualified. These levels represent different degrees of performance: no description of the skills associated with each of these levels is provided.[1] Boundary score zones between levels are set by reference to score statistics from past tests and evidence of how students in comparable college courses perform on the test. Test papers whose score falls at the boundary between two levels are examined to see how the examinee got that score: whether by doing very well on simpler items or by showing skills pertinent to the higher level.

[1]There are descriptive standards for performance on specific questions (to guide the scoring of those questions) but not for performance on the test as a whole.

19. NAGB comments that any conclusions about the reliability of the 1990 item judgments are "surmises" that no one has the data to disprove. The design of the 1990 item judgment process did indeed preclude a full assessment of the reliability of the judgment results. We quote NAGB's finding (included in its technical report on the project) that such evidence as was available suggested some problems.

20. NAGB states that we present no evidence that the procedure used to transform item judgment results to a NAEP score produced incorrect results. We agree that the procedure is accurate in this respect: if 48 percent is the standard, it locates a score on the NAEP scale at which students get 48 percent correct. Our question was whether students at this score got this percent correct by answering questions appropriate to their level. We have revised our presentation to make this more clear.

21. We are baffled by NAGB's comment that our analysis of NAEP results is flawed because it is based on "extrapolations rather than actual data." In fact, our analysis (see table 2.2) is based on data that NAGB itself presented to illustrate student performance at the 1990 achievement levels—data that NAGB describes as "the percentage of students in this group who gave the correct answer to the item."[2] Data computed in just the same way have been the basis for reporting student performance at various NAEP score levels for years and were still being used in 1992. Of course, the NAEP scores themselves are extrapolations (that is, statistical estimates), as explained in appendix III. But as we understand it, the figures in the table represent the actual performance of students whose estimated NAEP scores fell within the specified range.

22. Our analysis focuses on student performance on broad groups of items at different levels of difficulty, which is consistent with the mastery dimension of NAGB's achievement level definitions. (For example, we assume that students at the "advanced" score level should answer items in the most difficult group at a rate consistent with panelists' judgments.) We are not concerned with skill-by-skill specifications. But we are concerned that if NAGB purports to interpret NAEP scores in terms of item mastery (as in "partial mastery of fundamental skills"), the evidence should support this interpretation.

23. NAGB interprets the general agreement between judgment results and student performance as a confirmation of its procedures. We agree that

[2]National Assessment Governing Board, The Levels of Mathematics Achievement, vol. 2, State Results for Released Items (Washington, D.C.: 1991), p. 38.

performance at the basic score represents mastery of fewer and less difficult items than at the proficient level, which in turn represents mastery of fewer and less difficult items than at the advanced level. Our question was whether students at the scores selected met the mastery expectations (as well as the overall score expectations) for their level. The differences at the basic and advanced level seemed to us to be important from the point of view of the interpretations given to those scores, and especially to "below basic" scores.

24. NAGB perceives our report as being opposed to the use of descriptions of the kind of performance associated with a standard. Our position is that statements that describe expected performance at various levels are perfectly acceptable so long as they are demonstrably related to the actual performance exhibited at those levels. The problem with NAGB's descriptions is that this relationship has not been established.

25. We do not assert that the NAEP scale is flawed, nor do we argue that one cannot find percent correct equivalent scores on the NAEP scale. (See item 20.) We are concerned that the point thus selected may not represent the mastery expected for a given level—that the pattern of answers exhibited at a score of 255, for example, is not the pattern expected for students at the basic level. Our report both presents the logic from which our concern arose and provides evidence that the two patterns are different. NAGB asserts that this score is "most likely" to represent the basic level appropriately but does not support this assertion.

26. NAGB asserts that the conventional anchor points (which we describe as "straightforward" and "serviceable") raise the same issues as the achievement level descriptions (which we criticize). We see several significant differences between the two reporting methods. Anchor points are scores that are multiples of 50. Since the basis of selecting these scores does not reflect expectations of how students should perform, the issue of whether an anchor point represents the performance it is supposed to represent does not arise. Anchor point descriptions are indeed descriptions: they are based on observed performance at each anchor point score. NAGB's achievement level descriptions for 1990 combined prescription with description: they reflected a combination of expected and observed performance for each level. For 1992, the achievement levels descriptions state what panelists thought students should be able to do and do not reflect observed performance at all. To present them as if they describe actual performance would be misleading. Both the anchor level and the achievement level paragraphs are limited in that they focus on a

selected aspect of performance (that is, they are based on a limited set of items within a narrow range of difficulty) rather than on performance on a wide range of questions.

27. See item 8. We have deleted the words "reasonable and appropriate for the general student population of the United States" from chapter 3 since they apparently were subject to misinterpretation.

28. We present the GED example (minimum acceptable high school equivalency exam score set equal to the 30th percentile score earned by high school graduates) to illustrate a methodology. The use of this example does not imply either endorsement or criticism of the particular percentile selected. This methodology can be used to set standards at any level, from minimal to very challenging.

29. NAGB criticizes our comparison of U.S. and international achievement data and questions our calculations. Our presentation of international test data reflects published results; we made no calculations. We have deleted the comment characterizing the advanced standard as extreme.

30. NAGB mischaracterizes our example of how setting standards can be based on information about current performance. This comment represents a misconception that was also evident in NAGB's May 1990 achievement levels policy paper. Setting a percentile-based standard means selecting the test score earned by students at a given percentile in a base year, which could involve consulting a variety of evidence. For example, suppose NAGB concluded from various indicators that, in 1990, about 5 percent of American 8th graders were performing very well in mathematics. This would suggest that the 95th percentile score on the 1990 NAEP test, 317, might be an appropriate standard for advanced performance. Evidence of the NAEP performance of students whose estimated scores reached this level—and perhaps at the 90th percentile score as well—could be inspected. If performance at the higher (but not the lower) score appeared appropriately advanced, the score of 317 would be adopted as the standard. In 1992, scores of 317 or higher would be counted as advanced.

31. NAGB states that it "is not isolated from testing experts, as GAO suggests." This is not an accurate description of our position. We recognize that NAGB made increased use of technical resources after the problems with the 1990 achievement levels became evident. Our recommendations are aimed at ensuring that NAGB draws on such

resources as policy objectives are being formulated and delegates the implementation of technical procedures to appropriate experts.

32. NAGB argues that our recommendation that the achievement levels be withdrawn is unwarranted. Since we do not see evidence in NAGB's comments that would change our conclusions about the problems of interpreting the achievement level scores, we see no need to change our recommendation.

33. NAGB's statement that the achievement levels have "gone through" the 1992 prepublication review that has "dealt with . . . the question of . . . inferences" is true but does not tell the whole story. The review took the achievement level scores as given and included them in reporting NAEP results. However, the 1992 NAEP mathematics report notes that the question of what can be inferred from these scores has yet to be resolved.

34. NAGB states that the creator of the item judgment method, Dr. Angoff, found the achievement levels method suitable. Dr. Angoff was consulted in December 1989, before the specifics of the NAGB achievement levels approach had been developed.

35. NAGB terms our alternative methods of standard-setting "naive and unsupported by research evidence." The methods of setting overall performance standards we discuss in chapter 3 rest on informed collective judgments about what constitutes adequate performance on a test—judgments that can appropriately be made by a broadly representative governing body. Such judgments can be (and from what we have seen, typically are) informed by a variety of evidence, from performance-distribution data to item judgment results to inspection of student NAEP test papers and other work done by the students who produced these papers. As indicated in item 30, the simpler of these methods are not as naive as NAGB thought them.

36. NAGB again argues that we do not recognize that overall performance standards reflect expectations of student competence. In chapter 3, we make clear that judgments with respect to levels of overall performance reflect notions of what students should be able to do.

37. NAGB believes that we have overlooked evidence that it sought technical advice regarding the 1990 achievement levels. We have added to the text of chapter 4 to make clear that NAGB did seek and respond to expert advice with respect to the item judgment procedures and their

results after problems with its initial procedures had become evident. We also note that advice (including advice from NCES and ETS) was obtained throughout the process. Our concern is that NAGB did not respond to this advice, an observation that NAGB's comments do not dispute.

38. NAGB comments that the "policy on policies" does not apply to activities it implemented. We have amended the text of chapter 4 to respond to this comment.

39. NAGB argues that its overall performance is stronger than we suggest and that it has been successful in performing nontechnical functions. We have added a footnote to chapter 4 to make clear that our review covers only technical decisions and that our findings and conclusions do not reflect on NAGB's performance in other areas.

40 NAGB suggests that we draw too sharp a distinction between technical and policy issues. In fact, we recognize that most NAEP issues have both technical and policy aspects. Given that NAGB is not an expert body, we think its responsibility with respect to matters that are primarily technical should be only one of policy guidance. NAGB can appropriately raise policy issues and propose policy objectives, but it does not have the knowledge resources to prescribe or implement technical solutions.

41. NAGB argues that a strong, independent policy board is essential to ensure that constituency interests are represented and that it should not become a weak advisory committee, as we suggest. We do not suggest that NAGB should become a weak advisory committee; we simply suggest steps that could be taken to ensure that constituency interest is focused on the ends to be achieved through NAEP, leaving it to experts to determine whether and how those ends can be accomplished.

NAEP Summary Description

Statutory Purposes

NAEP's statutory purposes are to assess and to describe performance in the basic skills of reading, mathematics, science, writing, and history or geography on a regular schedule and in other subject areas as NAGB may direct. NAEP must present achievement fairly and accurately, based on test results from a representative sample of students, and must report data in a manner that supports valid, reliable measurement of trends.

Test Content

Traditionally, NAEP content has been designed to focus largely on the common ground of American education for each subject and grade—on the content that most students are taught. Tests have included relatively few items at the extremes of difficulty and relatively few that represent emerging practices. Beyond-grade items have been included in the 4th and 8th grade tests, however, in order to permit NAEP to form a single proficiency scale that cuts across grade levels (see below).

The skills and content to be covered by a NAEP test are identified through a broad-based consensus process, also required by statute, which involves teachers, curriculum specialists, local school administrators, parents, and concerned members of the general public. The consensus group sets the outline or framework for the test. The 1990 mathematics assessment framework called for coverage of five content areas: (1) numbers and operations; (2) measurement; (3) geometry; (4) data analysis, statistics, and probability; and (5) algebra and functions. Questions also covered three kinds of abilities: conceptual understanding, procedural knowledge, and problem solving. Content was drawn primarily from elementary and secondary school mathematics up to, but not including, calculus.

The test framework identifies the number of questions needed for each content and ability area at each grade level. A pool of items to implement that framework is designed, reviewed, and pretested. A given framework is used for several administrations of the test. Each time the test is given, some old items are dropped and some new ones are added. Eliminations and additions are made in such a way that the framework is not significantly altered, enabling comparisons to continue to be made over time.

Student Sampling

NAEP tests a sample of students from a sample of schools. The student sample is designed to be nationally representative. For the national 1990 mathematics assessment, the sample size was around 8,900 students per grade tested; about 6,400 usable responses per grade resulted.

Assignment of Items to Test Booklets

In order to reduce the testing burden for students, NAEP divides the items for an assessment into a number of blocks and assigns the blocks to test booklets according to a complicated statistical procedure. Each student receives one booklet and therefore sees only some of the items on the test. (For the 1990 8th grade mathematics test, each student saw around 59 out of 137 items.) The assignment of blocks of items and of students to test booklets is designed to ensure that nationally representative data are obtained for every test item. However, the sample of items that a given student sees is not necessarily representative of the test as a whole.

In addition to containing the test items, test booklets ask students for information about themselves and their parents and about the school's educational practices. Teachers and administrators also fill out questionnaires.

Scaling of Results

Since different students see different portions of the test, raw scores do not provide a comparable basis for reporting student performance on NAEP. NAEP uses complex statistical procedures to estimate how each student or someone of comparable background and proficiency would have performed if he or she had seen all the items on the test. The procedure takes into account the difficulty of the questions the student saw and answered, not simply the number of correct answers, and is designed to be empirically accurate. The statistically estimated scores, not the students' raw scores, are the basis for the NAEP reports.

Reporting

For all subjects other than writing, NAEP reports test results in terms of a 500-point scale that cuts across grade levels. The scale measures proficiency in the domain of elementary and secondary mathematics or other subject tested. Scores for 4th graders fall in the lower portion of the scale, scores for 8th graders in the middle portion, and scores for 12th graders toward the high end; there is overlap between grades.

To help readers understand the meaning of the scale in item mastery terms, NAEP summarizes and illustrates the types of items that students are first able to master consistently—items that reach the 65 to 80 percent correct level—at "anchor points" located at 50-point intervals along the scale (200, 250, 300, and so on). The item summaries are called anchor point descriptions. The percentage of scores that fall at or above each anchor point is also reported. However, until 1990 NAEP did not relate the anchor points to either content or performance standards.

Governance and Administrative Structure for the National Assessment

NAEP's governance structure includes two units: the National Assessment Governing Board and the National Center for Education Statistics.

NAGB Membership

2 governors or former governors

2 state legislators

2 chief state school officers

1 superintendent of a local educational agency

1 member of a state board of education

1 member of a local board of education

3 classroom teachers

2 curriculum specialists

2 testing and measurement experts

1 nonpublic school administrator or policymaker

2 school principals

1 representative of business or industry

3 representatives of the general public

NAGB Responsibilities

Formulating policy guidelines for NAEP

Selecting subject areas to be assessed

Identifying appropriate achievement goals

Developing assessment objectives

Developing test specifications

Designing the methodology of the assessment

Developing guidelines and standards for analysis plans and for reporting and disseminating results

Developing standards and procedures for interstate, regional, and national comparisons

Taking appropriate actions to improve the form and use of NAEP

NAGB also has final authority on the appropriateness of cognitive test items and is directed to ensure that such items are free from bias and that each learning area assessment has goal statements developed through a national consensus approach.

NCES Responsibilities

Carrying out NAEP with the advice of NAGB

Ensuring that NAEP provides a fair and accurate presentation of educational achievement, uses representative sampling, reports trends reliably, and includes information on special groups

Securing an independent evaluation of the Trial State Assessment

Ensuring the technical quality of the published data

Conducting reviews and validation studies of NAEP and soliciting comments on its conduct and usefulness

NAGB Achievement Level Descriptions: 1990 Mathematics

Fourth Grade

"BASIC: Partial Mastery of Knowledge and Skills. Fourth-grade students who are performing at the basic level should be able to solve routine one-step problems involving whole numbers with and without the use of a calculator. They should also be able to use physical materials and pictures to help them understand and explain mathematical concepts and procedures. Students at this level are beginning to develop estimation skills in measurement and number situations and should understand the meaning of whole number operations. For example, students performing at the basic level should be able to link the meaning of multiplication with the symbols needed to represent it. These students are also beginning to develop concepts related to fractions and read simple measurement instruments. Basic fourth-grade students should also be able to identify simple geometric figures and extend simple patterns involving geometric figures. These students should be able to read and use information from simple bar graphs."

"PROFICIENT: Solid Academic Performance. Fourth-grade students who are performing at the proficient level should have an understanding of numbers and their application to situations from students' daily lives. The proficient student should be able to solve a wide variety of mathematical problems; use patterns and relationships to analyze mathematical situations; relate physical materials, pictures, and diagrams to mathematical ideas; and find and use relevant information in problem solving. Fourth-grade proficient students should understand numbers and concepts of place value and have an understanding of whole number operations, as well as a facility with whole number computation. For example, students should be able to solve problems with a calculator and have the ability to use estimation skills to solve problems. Proficient fourth-grade students should understand and use measurement concepts such as length; be able to collect, interpret, and display data; and use simple measurement instruments."

"ADVANCED: Superior Performance. Fourth-grade students who are performing at the advanced level should be able to demonstrate flexibility in solving problems and relating knowledge to new situations. They should be able to use whole numbers to analyze more complex problems. Their understanding of fractions and decimals should extend to a number of representations. Students at this level should determine when estimation or calculator use is an appropriate solution to a problem, as well as read and interpret complex graphs. Advanced fourth-grade students should also be able to use measuring instruments in non-routine ways. These students should be able to solve simple problems involving geometric concepts and chance."

Eighth Grade

"BASIC: Partial Mastery of Knowledge and Skills. The eighth-grade student performing at the basic level should be able to identify and use the correct operations for solving one- and two-step problems involving addition, subtraction, multiplication, and division of whole numbers and decimals. These students should also have an understanding of place value and order of operations, and a conceptual understanding of fractions. They should be

able to use a calculator and estimation to arrive at answers to simple problems. Basic eighth-grade students can use rulers to calculate the perimeter and area of rectangular figures, and make conversions between units of measure within a given system of measurement. These students should be able to use basic geometric terms and identify elementary geometric figures. They should be able to read, interpret, and construct bar graphs and evaluate or solve simple linear equations involving whole numbers."

"PROFICIENT: Solid Academic Performance. Students at the proficient level should be able, with and without a calculator, to solve problems requiring decimals, fractions, and proportions. They should be able to compute with integers. They should be able to classify geometric figures based on their properties. Proficient eighth-grade students should be able to read, interpret, and construct line and circle graphs and show understanding of the basic concepts of probability. These students should be able to translate verbal problem situations into simple algebraic expressions and identify symbolic algebraic expressions representing linear situations."

"ADVANCED: Superior Performance. Eighth-grade students performing at the advanced level should be able to solve, with and without a calculator, a wide range of practical problems involving percents, proportions, and exponents. These students should have a solid conceptual understanding of the interrelationships among fractions, decimals, and percents and their connections with proportions. Eighth-grade advanced students should also understand and be able to use scale drawings, metric measurements, volume, and accuracy of measurement. These students should be able to solve problems involving elementary concepts of probability, interpret line graphs, and apply basic geometric properties related to triangles and to perpendicular and parallel lines.

Twelfth Grade

"BASIC: Partial Mastery of Knowledge and Skills. Twelfth-grade students who are performing at the basic level should demonstrate conceptual and procedural understanding of whole numbers, integers, fractions, and decimals and use them when solving routine problems. They should understand and apply measurement concepts and skills, including estimation, and solve routine problems involving time, money, and length. They should also be able to read scale drawings and use formulas to find areas and volumes. Basic twelfth-grade students should be able to identify a wide range of geometric figures, describe their characteristics, and solve problems involving angle measurements and similar triangles. These students should be able to interpret data in a variety of settings, including charts, tables, and graphs. Their understanding of chance should include the ability to select favorable outcomes to a situation and find the probability of an event in a setting involving a small number of outcomes. They should also be able to simplify and evaluate simple linear expressions and solve simple one-step linear equations and inequalities."

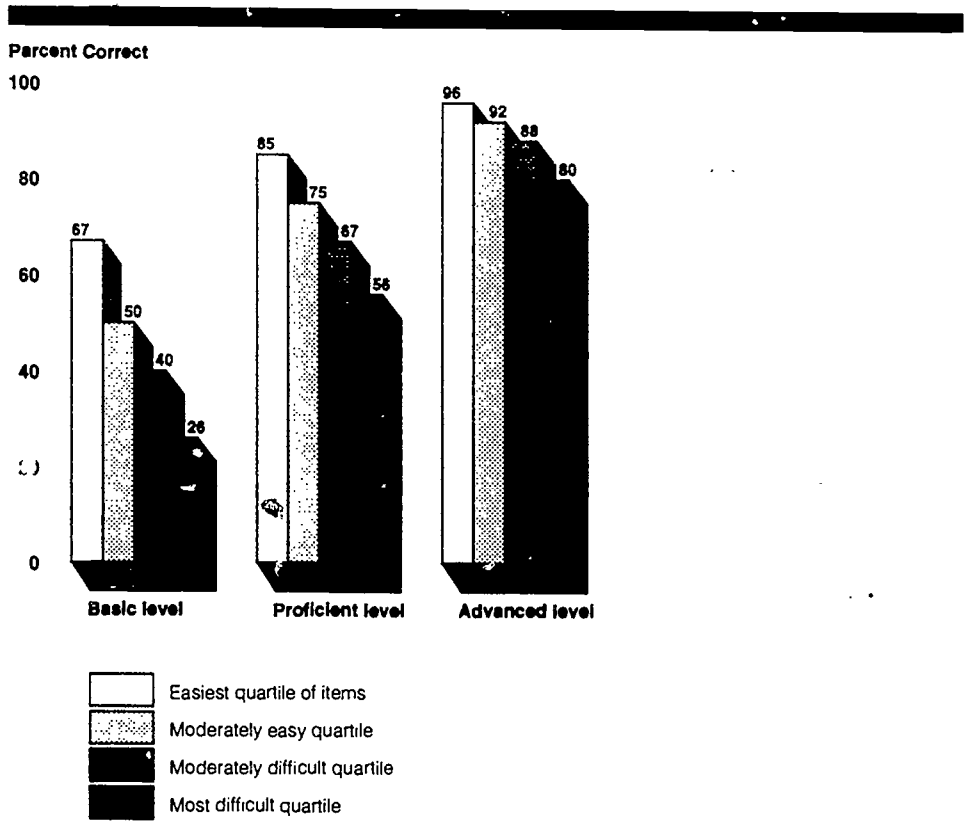
"PROFICIENT: Solid Academic Performance. Twelfth-grade students who are performing at the proficient level should have considerable command of the use of number and operations involving all forms of real numbers. In particular, these students should be able to represent problems involving integers, decimals, and fractions using symbols or graphs. These students should also be able to select, interpret, and use measurement relationships and formulas in problem situations. They should be able to make and evaluate conjectures about the properties of geometric figures. Proficient twelfth-grade students should be able to relate data about chance to physical models and use such models to solve problems. These students should be able to use coordinate systems on a number line to represent solutions to one-variable inequalities and use ordered pairs to describe locations in the plane."

"ADVANCED: Superior Performance. Twelfth-grade students who are performing at the advanced level should be able to investigate numerical relationships and determine the validity of conjectures involving number theory concepts such as parity (odd, even) and divisibility. These students should be able to establish procedures for the comparison and conversion of measurements of length, area, volume, and capacity. These students should understand the Pythagorean theorem and its applications, as well as use of coordinate geometry to represent relationships and solve problems. These students should also be able to graphically describe data for a situation, as well as provide numerical measures of central tendency (mean, median, and mode) and variability. Advanced twelfth-grade students should be able to apply probability and statistics concepts in reasoning about population characteristics based on information derived from a sample, including judging the adequacy of the sample. They should also be able to determine the probability of diverse events. These students should be able to translate information about linear situations from verbal or tabular forms to equations and analyze, verbally or in writing, the nature of relationships involving change in the values of the variables involved. These students should also be able to solve linear equations, inequalities, and systems of two equations in two variables, as well as evaluate a linear function and relate the value to a point on a graph of the function."

Calculation of Patterns of Performance

To calculate how NAGB's panelists judged that students should perform on items of varying difficulty (the performance pattern) as shown in table 2.2, we listed the 137 items on the 8th grade mathematics exam in order, from the least to the most difficult, based on judges' estimates of the percentage of students at the basic level who should get the item correct. (The item groupings for the proficient and advanced level were the same as for the basic level.) We then divided the items into four groups of equal size (34 items per quartile, except that there were 35 in the easiest quartile) and computed the average "percent correct" score for each group of items for basic, proficient, and advanced students according to panelists' item judgments. The results are shown in full in figure VI.1.

Figure VI.1: Eighth Grade Performance Pattern: Item Judgments[a]

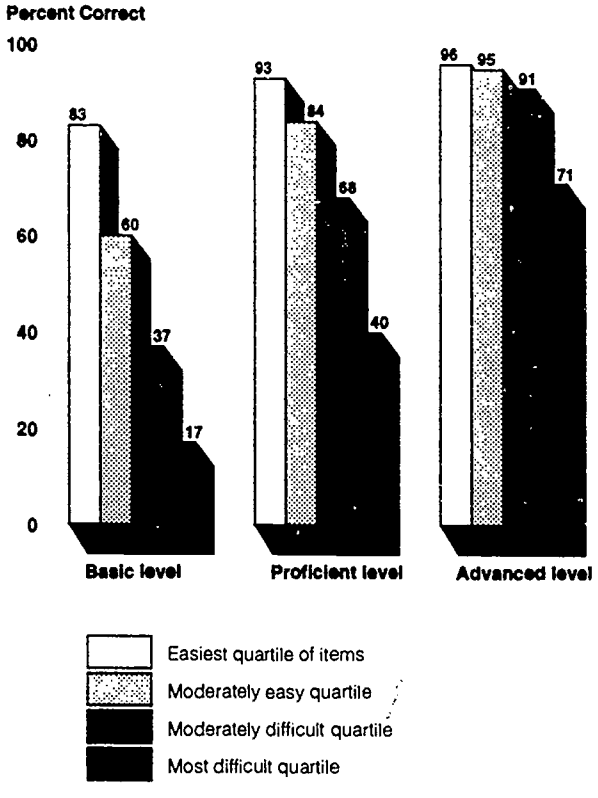


[a]The percent correct is the proportion of students at a given level who should answer items in each quartile correctly, according to panelists' item judgments.

Source: National Assessment Governing Board, The Levels of Mathematics Achievement, vol. 3, Technical Report (Washington, D.C.: 1991), pp. 265-71.

To obtain the measure of actual performance shown in table 2.2, we listed the 61 items that were made public from the 8th grade test in order of difficulty, as measured by the overall percent correct score reported for each item. We found the quartiles of difficulty for the full set of 8th grade items and assigned each of the published items to the proper item group. We then computed the average for each group in the same manner as for the item judgment data. The results are shown in figure VI.2.

Figure VI.2: Eighth Grade Performance Pattern at NAEP Score for Each Level[a]



[a]The percent correct is the percentage of students with a NAEP score 12.5 points below to 12.5 points above the standard for each level who answered items in each quartile correctly.

Source: National Assessment Governing Board, The Levels of Mathematics Achievement, vol. 2, State Results for Released Items (Washington, D.C.: 1991), pp. 5-35.

Since the item judgment data did not identify each test item by number, we do not know whether the items that fell into each group in the two sets of data were exactly the same or not. However, we know that the item

judgments were highly correlated with actual item difficulty—that is, that judges could distinguish easier from more difficult items. Thus, we expect that differences between the two sets of item groups are likely to be minor.

Major Contributors to This Report

Program Evaluation and Methodology Division

Frederick V. Mulhauser, Assistant Director
Gail S. MacColl, Project Manager
Venkareddy Chennareddy, Referencer
Penny Pickett, Reports Analyst

Bibliography

Standards for Educational Testing

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, D.C.: 1985.

Joint Committee on Testing Practices. Code of Fair Testing Practices in Education. Washington, D.C.: 1988.

Education Standards and Student Achievement: Reports

College Board. Academic Preparation for College. New York, N.Y.: 1983.

Educational Testing Service. Learning Mathematics. Report on the International Assessment of Educational Progress. Princeton, N.J.: 1992.

National Assessment Governing Board. The Levels of Mathematics Achievement. 3 vols. Washington, D.C.: 1991.

National Center for Education Statistics. NAEP 1992 Mathematics Report Card for the Nation and the States. Washington, D.C.: April 1992

National Center for Education Statistics. The State of Mathematics Achievement. Washington, D.C.: 1991.

National Center for Education Statistics. The Technical Report of NAEP's 1990 Trial State Assessment Program. Washington, D.C.: 1991.

National Council on Education Standards and Testing. Raising Standards for American Education. Washington, D.C.: 1992.

National Council of Teachers of Mathematics. Curriculum and Evaluation Standards for School Mathematics. Reston, Va.: 1989.

National Education Goals Panel. National Education Goals Report: Building a Nation of Learners, 1991. Washington, D.C.: 1991.

Office of Technology Assessment. Testing in American Schools: Asking the Right Questions. Washington, D.C.: U.S. Government Printing Office, February 1992.

Education Standards and Student Achievement: Articles

Angoff, William. "Scales, Norms, and Equivalent Scores." In R. L. Thorndike, ed. Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Berk, R. A. "A Consumer's Guide to Setting Performance Standards on Criterion Referenced Tests." Review of Educational Research, 56 (1986), 137-72.

Busch, John C., and R. M. Jaeger. "Influence of Type of Judge, Normative Information, and Discussion on Standards Recommended for the National Teacher Examination," Journal of Educational Measurement, 27:2 (1990), 145-63.

Cascio, Wayne F., R. A. Alexander, and G. V. Barrett. "Setting Cutoff Scores: Legal, Psychometric, and Professional Issues and Guidelines." Personnel Psychology, 41:1 (1988), 1-24.

Forsyth, Robert A. "The NAEP Proficiency Scales: Do They Yield Valid Criterion-Referenced Interpretations?" Educational Measurement: Issues and Practice, 10 (1991), 3-9 and 16.

Glass, G. V. "Standards and Criteria." Journal of Educational Measurement, 15 (1978), 237-61.

Hambleton, Ronald K. "Principles and Selected Applications of Item Response Theory." In Robert L. Linn, ed. Educational Measurement, 3rd ed. New York, N.Y.: American Council on Education and Macmillan Publishing Co., 1989. Pp. 147-200.

Hambleton, Ronald K., and Sally Powell. "A Framework for Viewing the Process of Standard Setting." Evaluation and the Health Professions, 6:1 (March 1983), 3-24.

Jaeger, Richard M. "Certification of Student Competence." In Robert L. Linn, ed. Educational Measurement, 3rd ed. New York, N.Y.: American Council on Education and Macmillan Publishing Co., 1989. Pp. 485-514.

Jaeger, Richard M. "Selection of Judges for Standard Setting." Educational Measurement: Issues and Practice, 10:2 (Summer 1991), 3-6.

Johnson, Eugene G. "Defining Levels on the 1990 Mathematics Composite." Presented at the annual meeting of the American Education Research Association, Chicago, Ill., 1991.

Kane, Michael, and Jennifer Wilson. "Errors of Measurement and Standard Setting in Mastery Testing." Applied Psychological Measurement, 8:1 (1984), 107-15.

Maurer, T. J., et al. "Methodological and Psychometric Issues in Setting Cutoff Scores Using the Angoff Method." Personnel Psychology, 44:2 (1991), 235-62.

Mills, Craig N., Gerald J. Mellican, and Nancy Thomas Ahluwalla. "Defining Minimal Competence." Educational Measurement: Issues and Practice, 10:2 (Summer 1991), 7-10.

Norcini, John J. "Equivalent Pass/Fail Decisions." Journal of Educational Measurement, 27:1 (1990), 59-66.

Norcini, John J., Judy A. Shea, and James C. Ping. "A Note on the Application of Multiple Matrix Sampling to Standard Setting." Journal of Educational Measurement, 25:2 (Summer 1988), 159-63.

Norcini, John J., et al. "The Effect of Various Factors on Standard Setting." Journal of Educational Measurement, 25:1 (Spring 1988), 57-65.

Petersen, Nancy S., Michael J. Kolen, and H. D. Hoover. "Scaling, Norming and Equating." In Robert L. Linn, ed. Educational Measurement, 3rd ed. New York, N.Y.: American Council on Education and Macmillan Publishing Co., 1989. Pp. 221-62.

Plake, Barbara S., and Michael T. Kane. "Comparison of Methods for Combining the Minimum Passing Levels for Individual Items into a Passing Score for a Test." Journal of Educational Measurement, 28:3 (1991), 249-56.

Reid, Jerry B. "Training Judges to Generate Standard-Setting Data." Educational Measurement: Issues and Practice, 10:2 (Summer 1991), 11-14.

Shepard, Lorrie A. "Evaluating Test Validity." In Linda Darling-Hammond, ed. Review of Research in Education, 19 (1993), 405-50.

Other Documents

American Council on Education. Examiner's Manual for the Tests of General Educational Development, 1991. Washington, D.C.: 1991.

Astin, A. W., et al. The American Freshman: National Norms for Fall 1991. Los Angeles: Higher Education Research Institute, UCLA, 1991.

College Board. Grading, Interpreting, and Using Advanced Placement Examinations. New York, N.Y.: College Entrance Examination Board, 1991.

Livingston, Samuel A., and Michael Zieky. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton, N.J.: Educational Testing Service, 1982.

New York State Education Department. Report to the Board of Regents on Career Preparation Validation Study. Albany, N.Y.: n.d.

U.S. Department of Labor, Secretary's Commission on Achieving Necessary Skills. Skills and Tasks for Jobs. Washington, D.C.: U.S. Government Printing Office, 1992.

Related GAO Products

Educational Testing: The Canadian Experience With Standards, Examinations, and Assessments (GAO/PEMD-93-11, April 28, 1993).

Planning for Education Standards (GAO/PEMD-93-21R, April 12, 1993).

Student Achievement Standards and Testing (GAO/PEMD-T-93-1, Feb. 18, 1993).

Student Testing: Current Extent and Expenditures, With Cost Estimates for a National Examination (GAO/PEMD-93-8, Jan. 13, 1993).

National Assessment Technical Quality (GAO/PEMD-92-22R, March 11, 1992).

Ordering Information

The first copy of each GAO report and testimony is free. Additional copies are $2 each. Orders should be sent to the following address, accompanied by a check or money order made out to the Superintendent of Documents, when necessary. Orders for 100 or more copies to be mailed to a single address are discounted 25 percent.

Orders by mail:

U.S. General Accounting Office
P.O. Box 6015
Gaithersburg, MD 20884-6015

or visit:

Room 1000
700 4th St. NW (corner of 4th and G Sts. NW)
U.S. General Accounting Office
Washington, DC

Orders may also be placed by calling (202) 512-6000 or by using fax number (301) 258-4066.

United States
General Accounting Office
Washington, D.C. 20548

Official Business
Penalty for Private Use $300

First-Class Mail
Postage & Fees Paid
GAO
Permit No. G100