Setting Performance Standards for Student Achievement

A Report of

the National Academy of Education Panel

on the Evaluation of the NAEP Trial State Assessment:

An Evaluation of the 1992 Achievement Levels

Principal Investigator
Lorrie Shepard, University of Colorado

Panel Chairmen
Robert Glaser, University of Pittsburgh
Robert Linn, University of Colorado

Project Director
George Bohrnstedt, American Institutes for Research

Printed on recycled paper in the United States of America
The National Academy of Education
Stanford University
School of Education CERAS-507
Stanford, CA 94305-3084
(415) 725-1003

The National Academy of Education

The National Academy of Education is composed of scholars and education leaders who "promote scholarly inquiry and discussion concerning the ends and means of education, in all its forms, in the United States and abroad." Our current active membership is limited to 125 scholars. The heart of the Academy is found in the serious and vital exchanges that take place in our regular meetings and the special commissions we establish. Throughout our 28-year history, the Academy has been called upon by governmental and other agencies to conduct special studies and reviews in the public interest. The most comprehensive evaluation of the National Assessment of Educational Progress ever conducted was organized by the Academy and directed by Lamar Alexander and H. Thomas James in 1986. We served as fiscal agent and coordinator of the Alexander-James Panel, and we provided a widely-cited independent review of that endeavor. Most recently, we completed a study, funded by the Carnegie Corporation of New York, to recommend priorities for the funding of educational research, entitled "Research and the Renewal of Education."

As the national trend toward education reform broadens and accelerates, we focus again on NAEP. This time we contribute the intellectual leadership and coordination for an independent evaluation of the 1990 and 1992 Trial State Assessments, which provide state-by-state comparisons of educational achievement, the first such use of NAEP data. Specifically, we are conducting a carefully constructed portfolio of studies using a range of methodological approaches. A panel of experts, selected by the Academy on the basis of their distinguished accomplishments, manages and coordinates the study portfolio. The Panel and the Academy examine a range of major issues: state-by-state and state-to-national comparisons; the validity and reliability of the assessment data; content, curriculum, and consensus processes; operations, sampling, data analysis; and the methods of reporting results of the assessments.

This report is focused specifically on the issues surrounding the definition, validity, and use of achievement levels in the 1992 NAEP Trial State Assessment. The Academy submitted the first of two mandated reports about the Trial Assessments in November of 1991. The second will be submitted in November of 1993. There will also be a capstone report drawing together information from various sources to document the quality and validity of the Trial State Assessment data.

Carl F. Kaestle
President, The National Academy of Education

Table of Contents

Detailed Table of Contents

July 30, 1993

Emerson Elliott
Commissioner
National Center for Education Statistics
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, DC 20208-5653

Dear Emerson:

On behalf of The National Academy of Education, I am pleased to transmit to you a report of the Academy's Panel on the Evaluation of the NAEP Trial State Assessment entitled *Setting Performance Standards for Student Achievement*. This report, requested by the National Center for Education Statistics, is focused on the specific issue of setting achievement levels in connection with the 1992 National Assessment of Educational Progress. The report discusses the definition, the validity, and the use of achievement levels such as those used in the 1992 assessments.

This report carries the approval of the Panel and has been reviewed and approved by The National Academy of Education's Executive Council, acting as a Committee of Readers. On behalf of the Executive Council of the Academy and its entire membership, I am pleased to present this report. Its findings and recommendations should assist policymakers in reaching thoughtful, informed decisions about this issue.

As the new President of The National Academy of Education, I have been impressed by the scholarly expertise and the searching discussions of this Panel. The Panel's chairs, Robert Glaser and Robert Linn, join me in expressing our hopes that NAEP will maintain its high standards and its integrity as a unique source of information about trends in the achievement of the nation's youth, a concern which we know you share.

Sincerely,

Carl F. Kaestle

Carl F. Kaestle
President, The National Academy of Education
William F. Vilas Professor of Educational Policy
Studies and History, University of Wisconsin

Acknowledgments

This report would not have been possible without the assistance of many colleagues and associates who unselfishly provided support to the Panel. The Panel extends special thanks to the National Center for Education Statistics (NCES) for its invitation to evaluate the achievement levels. At NCES, Emerson Elliott, Commissioner of Education Statistics, provided expert guidance and assistance to the Panel. Gary Phillips, Eugene Owen, Steve Gorman, Tongsoo Song, and Sharif Shakrani responded quickly and thoughtfully to requests for help and information. The responsiveness and helpfulness of these NCES staff greatly facilitated the Panel's investigations and kept the project moving on a fast track.

Special thanks are also due to the National Assessment Governing Board (NAGB). Roy Truby, Mary Lyn Bourque, and Dan Taylor generously provided the Panel with access to meetings, data, and materials pertaining to the Achievement-Level-Setting Process. The Panel also extends thanks to the many staff at American College Testing (ACT) who aided in this evaluation. Mel Webb and Susan Loomis (along with Howard Garrison of Aspen Systems) provided technical support and information to the many researchers who observed the Achievement-Level-Setting meetings all over the nation. Ric Luecht generously made achievement-level data available to Panel staff for re-analysis. The Panel is extremely grateful to NAGB and its contractors for their cooperation.

The assistance of the NAEP contractors, Educational Testing Service (ETS) and Westat Inc., and National Computer Systems, made the timely completion of several studies possible. Staff at ETS/NAEP who provided ongoing, long term support for multiple studies and to whom we owe a special debt of gratitude include Jules Goodison, Eugene Johnson, Ina Mullis, and John Olson. Other ETS/NAEP staff who contributed their expertise to different efforts include Claudia Gentile, Dave Freund, Bruce Kaplan, and Al Rogers. Nancy Caldwell at Westat greased the NAEP administrative wheels and helped make the 1991 Field-Test study a reality; other staff at Westat who assisted include Annelle Bond, Lesly Flemming, Cindy Randall, Sandy Rieder, and Dianne Walsh. In addition, Brad Thayer at National Computer Systems (NCS) provided invaluable support to the Field-Test study. Barry Druesne of the College Board (ETS), Walter MacDonald at the Advanced Placement program (ETS), and their respective staffs graciously fielded numerous phone and mail requests for unpublished statistics, information, and reports.

The Panel appreciates the excellent work and Herculean efforts put forth by the distinguished researchers who conducted the Achievement-Levels Evaluation investigations. Their careful study and detailed reports framed much of the Panel's thinking about the issues, and formed the backbone of this evaluation. Heartfelt thanks are extended to: P. David Pearson and Lizanne DeStefano (University of Illinois, Urbana-Champaign) for their studies of the Reading Achievement Levels; Ed Silver and Pat Kenney (University of Pittsburgh) for their studies of the Mathematics Achievement Levels; John Hayes (Carnegie-Mellon University) and Glynda Hull (University of California, Berkeley) for evaluating the Writing Achievement Levels; and Al Beaton and Eugenio Gonzalez (Boston College) for their study involving international comparisons.

Many, many persons helped carry out the data collection, information search, and other research activities necessary for this report. Although we cannot name each study participant separately in this report, we would like to thank the following individuals and groups for their kind and generous assistance: Thomas Boysen, Kentucky Commissioner of Education; Kevin Hill, Kentucky State Department of Education; Richard Hill at Advanced Systems and his staff; Wayne Martin, Colorado Department of Education and the EIAC Special Task Force on Assessment; and Jim Maxey at ACT. Special thanks go to Michael Kirst (Stanford University) and Tom Romberg (NCRMSE) for their willingness to read and review sections of this report.

The Panel thanks all of the teachers who participated in the 1991 NAEP Field-Test Special Study by furnishing ratings or hosting classroom visits, plus all the administrators and Westat field supervisors who helped. We deeply appreciate the help of all the northern California teachers who participated in the Palo Alto replication studies. Finally, we would like to thank all of the teachers, educators, and researchers who participated in the Mathematics, Reading, and Writing expert panel meetings.

Foreword

In the recounting of our nation's drive toward educational reform, the last decade of this century undoubtedly will be identified as a time when a concentrated press for national education standards emerged. The press for standards was evidenced by the efforts of federal and state legislators, presidential and gubernatorial candidates, teacher and subject-matter specialist councils, governmental agencies, and private foundations. There was little doubt of the importance of high goals and standards as catalysts for educational improvement. National standards for defining challenging content and performance expectations were seen as critical to the future of the nation—to address national needs for educational equity, to improve international competitiveness, to enrich students' capacities to use what they know outside of school, including lifelong learning, and to enhance the general culture.

A significant factor, emphasized over and over again, in striving to attain these goals was the necessity to assess students' levels of achievement—not only what they know, but also what they should accomplish.

In this context, the National Academy of Education Panel that was created to evaluate the Trial State Assessment component of the National Assessment of Educational Progress (NAEP) was charged with the added responsibility of evaluating the efforts to use NAEP to establish national performance standards. In its work, the Panel has found it essential to consider NAEP in the context of myriad policy and technical factors that can affect the validity and interpretation of the nation's key indicator of educational status and progress. Various criteria must be considered in judging the value of educational indicators including: *Credibility*—Standards should address the content and performance levels judged to be of greatest priority to the educational community, subject-matter specialists, and the community-at-large. *Public Understanding*—Standards should be understandable and relevant to a broad audience, including those who plan, manage, deliver, use, and pay for educational services. *Balance and Comprehensiveness*—The assessment of attained standards should be adequate for reporting both the outcomes of current school activity and of future achievement objectives, which suggest alternate assessments and new content standards for evaluating progress. *Measurability*—Objectives should be presented in terms of scales that lead to appropriate and accurate interpretations. *Policy Relevance*—Content and performance standards should reflect the concerns and engage the participation of professionals, advocates, and consumers, together with relevant federal, state, and local education departments.

The National Assessment Governing Board (NAGB) has, among its many responsibilities, confronted the important issues of the interpretability and public comprehension of NAEP results. In order to define what achievement ought to be, the Governing Board has set achievement levels for NAEP as standards that define what students in grades 4, 8, and 12 should know and be able to do. It is NAGB's intention that the achievement levels make NAEP far more useful for parents and policymakers by providing performance standards against which to measure educational progress and the attainment of the national education goals. By setting standards, they serve as an inducement to higher achievement. In NAGB's view, the use of achievement levels for reporting should increase the value of NAEP results to the American public.

In pursuing this task, NAGB identified a set of cutpoints on the NAEP scale to define basic, proficient, and advanced achievement levels at grades 4, 8, and 12, and proceeded to use these levels in reporting NAEP results in 1990 and 1992. There are many policy, procedural, and technical requirements for the development of achievement levels of this kind if they are to accomplish the objectives stated above. Shortcomings in the process, on the other hand, can result in inappropriate levels and misleading interpretations.

It is the process used, and the results of NAGB's approach to setting achievement levels, that the National Academy of Education Panel has been asked to evaluate. Toward this end, the analysis presented in this report is both detailed and broad, consisting of results of a wide array of studies of the process of deriving achievement levels and of the properties of the results obtained. The results of the detailed studies provide grounding for our broader deliberations about the setting of educational performance standards for the nation.

We are aware that our work has taken place in a period of heightened activity with respect to proposals for standards for student achievement and assessment. Numerous innovations are being implemented at the national, state, and local levels. We see NAEP as an extremely valuable independent monitor of the status and trends of change in student achievement as our nation proceeds toward improved environments for education and school reform. The observations in this report will hopefully contribute to this purpose. As educational reforms proceed in our country, it is the belief of the Academy Panel that systematic study of innovations and their results should continue to be an essential part of these efforts.

Robert Glaser, Chairman
Robert Linn, Co-Chairman

August 2, 1993

The National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment Project

Robert Glaser, Chairman
Director, Learning Research and Development Center
and National Research Center on Student Learning
University of Pittsburgh

Robert Linn, Chairman
Co-director, National Center for Research
on Evaluation, Standards, and Student Testing
University of Colorado at Boulder

Gordon Ambach
Executive Director
Council of Chief State School Officers

Isabel Beck
Professor of Education
Senior Scientist, LRDC
University of Pittsburgh

Lloyd Bond
Professor, Educational Research Methodology
University of North Carolina at Greensboro

Ann Brown
Professor of Education in Math, Science, and Technology
Evelyn Lois Corey Fellow in Instructional Science
University of California at Berkeley

Iris Carl
National Council of Teachers of Mathematics (NCTM)

Alonzo Crim
Professor of Education
Spelman College

Robert M. Groves
Professor of Sociology and Associate Director,
Joint Program in Survey Methodology
University of Michigan

Edward Haertel
Professor of Education
Stanford University

Lyle Jones
Professor of Psychology and Director,
L.L. Thurstone Psychometric Laboratory
University of North Carolina at Chapel Hill

Edward Roeber
Director of Student Assessment Programs
Council of Chief State School Officers

Albert Shanker
President
American Federation of Teachers

Lorrie Shepard
Professor of Education
University of Colorado at Boulder

Project Staff
American Institutes for Research

George Bohrnstedt, Project Director
Don McLaughlin, Associate Project Director
Shannon Daugherty
Phyllis DuBois
Dey Ehrlich
Jeremy Finn
Elizabeth Hartka
Susan Kleimann
Elise McCandless
Cathy O'Donnell
Frances Stancavage
Jean Wolman

Learning Research and Development Center

Mary Ann Thomas
Cindy Yockel

Executive Summary

Setting Performance Standards for Student Achievement

At the 1989 Education Summit in Charlottesville, Virginia, President Bush and the nation's governors agreed on six broad goals for education to be reached by the year 2000. Two of the goals pertained to student subject-matter achievement and had implications for assessment of educational progress—American students will leave grades 4, 8, and 12 with demonstrated competencies in challenging subject matter including English, mathematics, science, history, and geography; and U.S. students will be first in the world in science and mathematics by the year 2000. The National Education Goals Panel was then created and charged with measuring progress toward the goals developed at the Education Summit. The Goals Panel took on the challenge of determining how progress toward meeting national goals might be measured in the various subject-matter fields. The creation of national education goals led to the question, unprecedented in the nation's history, of whether there should also be national education standards.

Congress pursued the issue of establishing national standards by creating the National Council on Education Standards and Testing (NCEST) in June 1991. NCEST delivered a report, *Raising Standards for American Education,* to Congress in January 1992 that endorsed the development of education standards.[1] In developing recommendations on education standards, NCEST drew a useful distinction between content standards and performance standards: "Content standards describe the knowledge, skills, and other understandings that schools should teach in order for students to attain high levels of competency in challenging subject matter; performance standards define various levels of competence in the challenging subject matter set out in the content standards."[2]

The momentum to establish national standards continues to the present. In April 1993 President Clinton transmitted the *Goals 2000: Educate America Act* to the Congress. One of the major pillars of *Goals 2000* is to "develop voluntary academic standards and assessments that are meaningful, challenging, and appropriate for all students through the National Education Standards and Improvement Council." Additionally, various national content standards are being developed in several disciplines under the auspices of the U.S. Department of Education.

[1] National Council on Education Standards and Testing, *Raising Standards for American Education: A Report to Congress, the Secretary of Education, the National Education Goals Panel, and the American People* (Washington, D.C.: U.S. Government Printing Office, January 24, 1992, ISBN 0-16-036087-8).

[2] Ibid., 13.

NAEP as an Independent Monitor of the Status of Education

The National Assessment of Educational Progress (NAEP), first administered in 1969, is conducted under the auspices of the National Center for Education Statistics (NCES). NAEP, which is administered every 2 years, provides the best available trend information on the educational achievement of American students. The quality of NAEP rests, first, on the fact that its instruments are reasonably comprehensive measures of achievement in the core subject areas including reading, writing, mathematics, science, history, and geography. Second, NAEP assessments are administered to a probability sample of the nation's 4th, 8th, and 12th graders, thus ensuring that the results are representative of what U.S. students know at any given time. In its first report, *Assessing Student Achievement in the States,* this Panel recommended that "NAEP should be preserved as the nation's key independent indicator of how students are performing academically. NAEP has played and should play the same role for education that economic indicators have played in signaling changes in the status of the economy."[3]

The Creation of the National Assessment Governing Board and Its Decision to Develop Achievement Levels on the National Assessment

A year before the Education Summit, Congress reauthorized the National Assessment of Educational Progress. The law also created a National Assessment Governing Board (NAGB) to develop and oversee policy for the National Assessment. Prior to 1988, the law provided that the NAEP contractor appoint a separate Assessment Policy Committee to establish policy for each assessment. The legislation that reauthorized NAEP in 1988 charged NAGB with several responsibilities including "identifying appropriate achievement goals for each age and grade in each subject area to be tested under the National Assessment."[4]

The language in the 1988 reauthorization is similar to previous authorization language dating back to 1978 when the Assessment Policy Committee was given responsibility for "the design of the National Assessment, including the development and selection of the goal statements and assessment items... ."[5] NAGB might have responded to the charge in different ways. Given the emerging consensus for establishing national education standards, the fact that the Education Summit was silent on who should set standards, and the fact that NAEP was the only national assessment of achievement based on defensible samples, NAGB interpreted the authorizing legislation as a mandate to set performance standards, which it named "achievement levels," for

[3] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 67.

[4] Public Law 100-297, Part C, Section 3403 (6) (A): 1988.

[5] Public Law 95-561, Section 1242: November 1, 1978.

NAEP. In making this decision, the Governing Board took the first of a series of steps in moving from reporting what the nation's 4th, 8th, and 12th graders *actually* know, to reporting what they *should* know.

In addition to undertaking the development of achievement levels in response to the call for the establishment of national education goals, the Governing Board also saw achievement levels as a way to make NAEP results more understandable and hence more useful for advancing education reform in the nation. An early NAGB staff paper suggested that reporting results in terms of the proportion of students who attain substantive standards "may serve as a lever for school improvement."[6]

Standard-Setting Methods

A large number of standard-setting methods are documented in the tests and measurement literature. Judgmental methods require judges to set cutpoints after closely studying the items on a test, while empirical methods require judges to evaluate examinee performance outside of the test context and set cutscores based on the actual test performance of examinees they have judged to be competent. One consistent finding in the literature is that different standard-setting methods produce different results. In some cases, these methods even appear to be sensitive to slight and seemingly trivial differences in procedures.

The Angoff Method Chosen to Set Achievement Levels on NAEP

As a first step toward setting achievement levels on NAEP, the Governing Board selected the Angoff method, the most widely used and straightforward of the judgmental methods. They rejected the use of the contrasting-groups method, one of the best known empirical methods, because it would have required the administration of items to a trial population and thus delayed the immediate implementation of standards.

The Governing Board began by adopting generic definitions of basic, proficient, and advanced levels, presented in figure 1. These generic descriptors then served as the basis for more specific, substantive descriptions of the basic, proficient, and advanced levels in each subject area and at each of the three grade levels. NAEP scale cutpoints representing marginally basic, marginally proficient, and marginally advanced performance at each grade level were subsequently determined using the Angoff method. These cutpoints defined four score groups for each grade, the lowest being "below basic."

[6] Roy Truby, "The Future of NAEP as the Nation's Report Card" (Washington, D.C.: National Assessment Governing Board, November 29, 1989), 10.

Figure 1. ***Achievement-level definitions adopted by NAGB on May 11, 1990***

Basic

Denotes partial mastery of the knowledge and skills that are fundamental for proficient work at each grade—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

Represents solid academic performance for each grade tested—4, 8, and 12—and reflects a consensus that students reaching such a level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. At grade 12 the proficient level will encompass a body of subject-matter knowledge and analytical skills, of cultural literacy and insight, that all high school graduates should have for democratic citizenship, responsible adulthood, and productive work.

Advanced

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.

SOURCE: National Assessment Governing Board, *Achievement Level Options for the NAEP Mathematics Assessment: 1990 Trial* (Washington, D.C.: May 10, 1991).

Previous Evaluations Highly Critical

Because the achievement levels are the first highly visible examples of national performance standards, the Governing Board's standard-setting project has been evaluated more than any other aspect of NAEP. The most notable of the previous evaluations include studies by the NCES Technical Review Panel (TRP),[7] the Governing Board's own evaluators,[8] and the U.S. General Accounting Office (GAO).[9] The first two focused on the 1990 achievement levels while the GAO report spans both the 1990 and 1992 level-setting efforts.

[7] R.L. Linn, D.M. Koretz, E.L. Baker, and L. Burstein, *The Validity and Credibility of the Achievement Levels for the 1990 National Assessment of Educational Progress in Mathematics* (Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, January 1991).

[8] D. Stufflebeam, R.M. Jaeger, and M. Scriven, *Summative Evaluation of the National Assessment Governing Board's Inaugural 1990-91 Effort to Set Achievement Levels on the National Assessment of Educational Progress* (Washington, D.C.: National Assessment Governing Board, August 1991).

[9] U.S. General Accounting Office, *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations*. Report No. GAO/PEMD-93-12 (Washington, D.C.: Government Printing Office, June 1993).

The previous evaluations made several major criticisms. For example, the judgment tasks required by the modified Angoff process were found to be difficult and confusing; the NAEP item pool was not adequate to reliably estimate performance at the advanced levels; the standards set seemed highly dependent on the particular sample of judges; appropriate validity evidence for the cutscores was lacking; and neither the descriptions of student competencies nor the exemplar items were appropriate for describing actual student performance at the designated achievement-level cutscores. *All the evaluation studies concurred that the achievement levels, as constructed, were not appropriate for reporting NAEP results.*

The Governing Board was responsive to many of the concerns of its evaluators, and it did designate the 1990 achievement levels as a trial effort. However, NAGB remained committed to delivering final achievement levels for use in reporting 1992 results, and, consequently, advice that suggested the need for significant additional data collection or a fundamental rethinking of the achievement-level-setting process was not followed.

Despite reservations, NCES decided to make the 1992 results available in a timely manner by releasing the full state and national mathematics reports, with explanatory caveats, in April 1993. For the 1992 reading reports, scheduled for release in mid-September 1993, the achievement-level descriptions have been supplemented with empirically derived descriptions of skills and behaviors exhibited by students whose performance places them at a particular achievement level.

The National Academy of Education Panel's Evaluation

Because of the great interest in education standards and because of controversy surrounding earlier evaluations of the 1990 achievement levels, in 1992, NCES asked the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment to expand its work to provide an extensive evaluation of the 1992 achievement levels and the issues surrounding their use. The chair of NAGB's Achievement Levels Committee welcomed the Panel's observation and evaluation of the level-setting process, and draft legislation for the reauthorization of NAEP stipulated that the achievement levels be evaluated by the Panel.

Importantly, the Panel's report of the 1992 standard-setting effort is based on new evidence entirely independent of results of earlier evaluations. To the extent that the conclusions of past evaluations are similar to those of the present evaluation, the earlier reports lend additional weight to conclusions in this report.

Two themes are critically important in defining the policy context for this evaluation. First, the establishment of achievement levels for NAEP must be viewed in relation to the national effort to set education standards. Second, the achievement levels must be evaluated for their impact on the role of NAEP as an independent monitor.

Key Findings from the Panel's Evaluation Studies

The evaluation's overarching research question was whether use of the 1992 achievement levels would lead to valid interpretations of NAEP results. The Panel commissioned a series of in-depth studies to address: (a) the adequacy of the process used to develop the achievement levels; (b) the reasonableness of the level cutscores compared to external criteria; and (c) the validity of the achievement-level descriptions and exemplar items.

The Panel's Analyses of the Achievement-Level-Setting Process

The process of developing achievement levels involved two distinct tasks: (1) creating subject-specific descriptions for each level, and (2) identifying cutscores. In both reading and mathematics, the "initial" achievement-level descriptions created by the participants in the level-setting meetings were judged to be inadequate by subject-matter specialists and were substantially revised at a later date. The revisions caused a serious validity problem, however, because the achievement-level cutscores were never reset to correspond to the new descriptions.

The reading process evaluation documented one of the reasons for the inadequacy of the initial descriptions. Panelists were unfamiliar with the NAEP Reading Framework and therefore used personal experience and opinions to develop the descriptions and to make item judgments rather than following the framework.

The process used to set the 1992 cutscores in reading and mathematics was judged to be indefensible because of the large internal inconsistencies in judges' ratings. The Panel's analyses showed that judges could not maintain a consistent view of what a student at the borderline of each achievement level should be able to do. In some cases the internal inconsistencies were huge, with judges setting cutscores for the same level that differed by the equivalent of four to eight grade levels simply as a function of considering different item types. The modified Angoff process also did not facilitate the development of consensus. Differences among judges' ratings were large even at the end of a three-round process. *Based on its analyses, the Panel concludes that the Angoff procedure is fundamentally flawed for the setting of achievement levels.*

The Panel's External Comparison Studies

The weight of evidence suggests that the 1992 achievement levels were set unreasonably high. The Panel conducted four field-based studies (in reading and mathematics at grades 4 and 8) each involving over 1000 students. In three of the four studies, teachers' ratings and individual assessments administered by researchers consistently found more students performing at the advanced, proficient, and basic levels than were identified in these categories by the achievement-level cutscores. At grade 12, data from the Scholastic Aptitude Test and Advanced Placement examinations again

suggest that there are more advanced students in reading and mathematics than were found to be advanced by the achievement levels. The possibility that the cutscores are systematically too high is consistent with the finding from the Panel's content-expert studies in reading and mathematics, which showed that because there were no advanced items to measure the content of the descriptions, the experts moved higher and higher on the score scale in search of such items.

International data from the International Assessment of Educational Progress (IAEP) for 13-year-old students were analyzed in conjunction with the eighth-grade mathematics data from NAEP. However, IAEP data could not be used to evaluate the specific achievement-level cutscores because different countries and different percentile choices would each imply different cutscores. The Panel concluded that international data are useful in their own right for interpreting the relative performance of U.S. students, and the Panel suggests a means for establishing international benchmarks on the NAEP scale in its recommendations below.

The Panel's Studies on the Achievement-Level Descriptions and Exemplar Items

The validity of the achievement-level descriptions and exemplar items was evaluated by considering their congruence with emerging national content standards, the NAEP frameworks, and the NAEP item pools.

Current NAEP item pools, particularly at the advanced level, are not sufficiently congruent with emerging national content standards. Therefore, the achievement-level descriptions cannot adequately represent ideal future-oriented standards without departing from the assessment that students actually took. In addition, some exemplar items were judged by content experts to be less than exemplary. They do not communicate subject-matter standards well. Exemplar items shape public understanding about the nature of subject-matter expectations. If most members of the public are accustomed to viewing mathematics as number facts and algorithms, for example, then seeing such a narrow set of exemplar items will do little to convey higher expectations about mathematical thinking and new content standards.

Achievement levels are one kind of performance standard. They are intended to establish expectations for what students *should* know and be able to do in order to attain each level. In this sense, the standards are statements of desired rather than actual outcomes. However, once these idealized expectations have been articulated and used to specify cutpoints on the score scale, then it is understood that students who reach the level *can* do what is described at that level. Problems have arisen with the initial efforts to report NAEP results by achievement levels because in many instances it is not true that students at a level can do what is described by the level. Because of the ambiguity about the fit of the descriptions and exemplar items to the so-called advanced, proficient, and basic regions of the score scale, the 1992 report of NAEP mathematics results carried a number of caveats, which unfortunately obscured rather than illuminated the meaning of the results. The Governing Board and NCES significantly remedied one aspect of the "should versus can" problem for the 1992 reading assessment by anchoring the achievement levels with items that students at the various levels could, with high probability, do. While this change is an important

improvement for making accurate interpretations of what students at various levels of achievement can do, the change does not address the problems of the mismatch between the descriptions and cutpoints or the mismatch between elements of the descriptions and the content of the assessment.

The Panel concludes that flawed achievement levels would not enhance the interpretability of NAEP and might, in fact, jeopardize other national efforts to develop content and performance standards and might harm the credibility of NAEP.

The Panel's Recommendations

The members of the Panel strongly affirm the potential value of voluntary national standards that exemplify challenging curricular and performance expectations. However, the standards set must be defensible in order to ensure that assessment data and national education policy based on the standards are sound. Given the problems noted above, the Panel does not believe that the process by which the 1990 and 1992 achievement levels were set can be defended. In the Panel's judgment, setting credible performance standards is a long term process—standards cannot be set in 3 days nor in 3 months.

To deal with the tension between the need for time to develop long term performance standards that reflect emerging educational content standards on the one hand, and the desire for standards-based reporting to continue on the other hand, the Panel makes both short and long term recommendations.

Short Term Recommendations Regarding the Setting of Achievement Levels

1. **Discontinue Use of the Angoff Method.** The Panel recommends that use of the Angoff method or any other item-judgment method to set achievement levels be discontinued. As the Panel's studies demonstrate, the Angoff method approach and other item-judgment methods are fundamentally flawed. Minor improvements, such as allowing more discussion time or providing instructions about guessing, cannot overcome the nearly impossible cognitive task of estimating the probability that a hypothetical student at the boundary of a given achievement level will get a particular item correct. Furthermore, the Angoff method does not allow for an integrated conception of subject-matter proficiency.[10]

[10] Angoff procedures may or may not be defensible in other contexts (e.g., setting minimum standards based on all-or-none judgments about essential knowledge for a specific vocation). Based on the Panel's findings concerning internal and external validity, the NAE Panel is understandably skeptical. However, the Panel would not want the discussion here to be taken in any way as an indictment of other standards established by such methods.

2. **Discontinue Reporting by Achievement Levels as Used in 1992.** The Panel urges NCES and NAGB not to report the 1992 NAEP results by achievement levels. The descriptions and exemplar tasks are not adequate for reporting to the public what students should be able to do. Furthermore, the assessment content does not measure up to that expected for the emerging national content standards. Cutpoints set in both 1990 and 1992 warn us that attempts to establish performance standards with inadequate content run the risk of setting high levels of performance on low-level content. NAEP content must be expanded substantially to reflect emerging national content standards. When the content changes, trends based on the old levels will be potentially misleading. Thus it only makes sense to wait until national content standards are available and then to follow a more coherent process for developing performance standards in conjunction with content standards.

   The Panel is aware that plans for the NAEP assessments in 1994 are well underway. Indeed, the Governing Board has issued a request for proposals for the development of achievement levels for the U.S. history and geography assessments. The Panel recommends reconsideration of the intent to develop achievement levels, particularly with the use of the Angoff or any other item-judgment method for the 1994 assessment. The Panel also recommends against the use of the 1992 achievement levels in mathematics and reading as baselines against which to make comparisons in future assessments, given the flaws found with the Angoff process for the setting of achievement levels.

3. **Invite Content Experts, Business Leaders, and Standards Committees to Comment on the Meaning of NAEP Results and Desired Performance Standards.** The Panel recommends that beginning with the 1992 reading assessment, NAGB contribute to the dialogue about national content and performance standards by inviting different groups such as subject-matter experts and business leaders to study NAEP results closely and produce an evaluative commentary. Such groups could even be encouraged to consider the question "How good is good enough?" By adding policy interpretations after the reports are released to the public, rather than building them into the score reporting, it becomes possible to have a more informed debate about competing perspectives.

   Although the development of content standards should precede the development of performance standards, content standards cannot be developed purely in the abstract without concrete examples of instructional activities and student work. Therefore, the Panel specifically recommends that the Governing Board invite national standards committees in each subject area, such as the National Academy of Sciences/National Research Council science content committee, to examine NAEP results closely, at

the level of specific tasks and items, and report on their meaning. Exemplar items could be used to illustrate desired performance at various levels. The Panel expects the benefits of such an exchange to accrue on both sides. NAGB would gain an evaluative interpretation useful in the description of NAEP results; the standards group would gain insight into the measurement of performance standards. Finally, content-standards committees would also benefit from access to broadly representative examples of student performance and (indirectly) evidence of current teaching practices.

4. **Publish Achievement Levels in 1994 Separately from the Official NAEP Reports and Report These as Draft or Developmental.** The primary means of reporting in 1994 should be by average NAEP scale scores and anchor points. As a first step in the development of performance standards following the sustained and iterative process outlined in the section on long term recommendations below, the Panel recommends that efforts to establish experimental or trial achievement levels in reading, history, and geography begin in 1994. New standard-setting procedures should be developed and implemented on a trial basis. Before attempts are made to set cutscores for the 1994 assessments, descriptions, consistent with the NAEP frameworks, should be agreed upon. Rather than rely on a single procedure, potential achievement-level cutscores should be developed using several approaches, with followup efforts then made to reconcile differences and arrive at defensible standards. Based on insights from its evaluation studies, the Panel recommends that at a minimum *all three* of the following approaches be used: (1) contrasting-groups field-based studies, (2) an item-mapping procedure, and (3) a total-student-performance (whole-booklet) mapping procedure. As part of the process, exemplar items should be identified that meet both conceptual and statistical criteria.

   Efforts to set achievement levels using these new strategies must be carefully evaluated, and it is essential that the results be viewed only as a research and development effort. Therefore, the achievement-level results should be released by NCES as part of its ongoing Research and Development series of publications, preferably sometime after the release of the official NAEP results. To avoid previous problems that made it necessary to revise the achievement levels after the fact and recompute trend data, achievement levels for each subject area should be reported in draft form for at least one assessment cycle before being adopted as the official levels. Such a process, even in the short term, would allow adequate time to respond to criticism and evaluation findings.

Achievement levels should not become the primary method of reporting NAEP results until they have been through the substantive and empirical evaluations suggested in the long term recommendations. However, examples of types of benchmarks that have interpretive value and might be developed in the short run are suggested in sections below.

5. **Use 1990 and 1992 Percentile Scores to Monitor Achievement in Future Assessments.** Reporting NAEP results using performance standards closely aligned with national content standards is desirable, but the content standards are not yet established. In the meantime, ways of reporting NAEP are needed to help the public and policymakers (1) see trends in performance and (2) understand changes in various levels of performance. Therefore, the Panel recommends that baseline performance could be set at three levels using thoughtfully chosen percentile scores. NAGB, in consultation with NCES, should decide the three percentile cutscores. The Panel suggests the 95th, 75th, and 25th percentiles for a base year could be used as benchmarks against which to measure future progress. This procedure could be implemented on the 1990 NAEP mathematics assessment. Doing so would allow the Governing Board to trace progress in mathematics from 1990 to 1992 and then again from 1992 to 1995, or whenever the next mathematics assessment is administered. The Governing Board would decide what percentage of students should reach the three target levels in future assessments, and NCES would report progress against meeting those targets. For example, in setting a target for the year 2000, the Governing Board might decide that 20 percent of the students should be above the 1990 95th percentile benchmark in mathematics. As an alternative or addition, comparisons using international data (see recommendation 6 below) could be used.

   Baseline percentile scores can also be set for the 1992 reading assessment, allowing NAGB to set targets for the 1994 reading assessment. Similarly, baseline performance in history and geography, as well as target performances for future assessments in these content areas, can be determined after the 1994 assessment is administered and scored.

6. **Use International Comparisons to Set Benchmarks for U.S. Performance.** Comparisons between the mathematics achievement of U.S. eighth graders and the achievement of students from other nations can provide useful information. Therefore, in addition to using percentile scores on NAEP, the Panel recommends that NAGB and NCES compare the performance of U.S. students with that of students in other nations. While the details for making these international comparisons should be determined by NCES, one suggestion is to use international percentile results as benchmarks if defensible percentile equivalences can be established between NAEP and international assessments. For example, the score corresponding to the 90th percentile for the combined results of, say, the four highest scoring countries (e.g., on the 1991 International

Assessment of Educational Progress or the upcoming Third International Mathematics and Science study) could be identified as advanced performance, and the corresponding percentile rank for U.S. students could then be determined. The NAEP score corresponding to the latter percentile rank could then be used as a benchmark to track the progress of U.S. students on future NAEP assessments. That is, the percentage of students scoring at the benchmark level or above becomes the baseline against which to make future comparisons. It would be within the purview of the Governing Board to set target percentages to be reached in future assessments as goals towards which to strive. Stated differently, NAGB could target the percentage of students that it believes should be "world class" in mathematics and science by the year 2000.

7. **Work with the National Education Goals Panel to Develop a Way to Use NAEP Results to Measure Progress over the Decade of the 1990s.** The Panel recommends that NAGB and NCES join with the National Education Goals Panel to work out a mutually acceptable way for reporting the results over the decade of the 1990s (i.e., until such time as performance standards based on certified national content standards can be developed). The Panel has suggested several possible strategies for reporting NAEP results in the short run. NAGB, NCES, and the Goals Panel should agree on a common method for reporting progress. A common method is important because, as the Panel noted in its evaluation of the 1990 Trial State NAEP, it was confusing for the public to have the same results reported in different ways by NCES and the Goals Panel.

8. **Implement Within-Grade Score Reporting.** At present, performance for students in grades 4, 8, and 12 is reported on a single 0-to-500-point scale. As part of its argument for achievement levels, NAGB expressed concern that use of a single scale obscured variation within grade and made it more difficult to make evaluative judgments.

   The Panel considers the use of a single vertical scale to be problematic on other grounds. The decision to have a subject-matter scale span multiple grades complicates the meaning of a composite score because weighting of the composite changes with grade level. It is uncertain, for example, whether a score of 250 has the same meaning for a fourth and eighth grader.

   During the course of the evaluation studies, the Panel heard many comments from subject-matter experts proclaiming that a strength of the achievement levels was their focus on describing performance within grade level. In the absence of achievement levels, exemplar items based on selected within-grade anchor points could be used to describe what performance at various levels looks like. The Panel also believes that within-grade scales will make it easier for subject-matter experts to comment meaningfully on the quality of NAEP performances. For these reasons, the Panel recommends moving to within-grade scaling effective with the 1994 assessment.

The eight recommendations detailed in the section above are short term in the sense that the Panel believes that they can and should be implemented in the near future. Importantly, the Panel believes that these recommendations should remain in place into the long term as well, with the possible exception of recommendation 5 regarding percentile scores which may be superseded by performance standards of the type described below.

Long Term Recommendations for Developing National Performance Standards Congruent with National Content Standards

As national content standards are developed and certified, the Panel believes it is imperative that performance standards on NAEP be linked to them. This will be a time-consuming process. The Panel also believes that the development of such performance standards requires a knowledge base for understanding the meaning of various levels of performance. A knowledge base of this sort cannot be developed quickly enough to be available for the next assessment cycle. For these reasons, the Panel believes that the Governing Board must also take a long view as it seeks to establish performance standards. With this perspective in mind, we turn to the Panel's long term recommendations.

1. **Develop Content Standards and Performance Standards in an Iterative Process.** Because performance standards must include descriptions of what students at each level should be able to do, along with a quantitative score defining that level, the Panel recommends that performance standards be developed in close coordination with emerging national content standards and assessment tasks. As these content standards become available in each discipline, performance standards can then be created consistent with those content goals. Answering the question of how much students should know without first establishing content standards is virtually impossible. Therefore, measuring progress toward national goals in all NAEP subject areas must await the development of additional national content standards.

2. **Establish a Standing Subject-Matter Panel for Each Subject Area.** In its last report, the Panel argued that fragmentation of assessment development could be alleviated if subject-matter panels were established with ongoing responsibility for overseeing all phases of the assessment. This same idea is even more important in the context of developing performance standards. In keeping with a consensus approach, the Panel recommends the establishment of standing subject-matter panels that are broadly representative of classroom teachers, curriculum experts, members of the lay public with expertise in the subject area, and researchers. The standing panel in each subject should participate in the development of new frameworks. As part of their deliberations, the standing subject-matter panel should consider explicitly how to incorporate national standards, how to

reflect current practice, and whether these perspectives should or should not be combined in a composite score. The same panel would review item development, agree on narrative descriptions of performance standards, and work on selecting representative tasks until there was both logical and statistical correspondence between content standards, performance standards, and illustrative tasks.

3. **Address Important Conceptual Issues.** The Panel recommends that subject-matter panels working with teams of developmental specialists be asked to address, over a period of time, conceptual issues that have thus far not been considered as part of the effort to specify achievement levels. These issues have to do with underlying assumptions regarding the nature of proficient and advanced performance in each subject area and the development of proficiency across grade levels. To date, there has been no theoretical or developmental model underlying NAEP scales. Rather, the scales have been created statistically, without evaluating the implied developmental model that underlies them.

4. **Empirically Evaluate Achievement Levels Before Making Them Operational.** After subject-matter panels have developed performance standards for NAEP, the Panel recommends that the standards be rigorously evaluated before making them an operational part of the reporting of NAEP results. In particular, the Panel recommends that the levels be validated using the contrasting-groups method that was used to evaluate the current set of achievement levels. Performances should be examined to ensure that the observed patterns confirm predictions based on the explicit or implicit conceptual models used to formulate the performance standards.

5. **Recognize the Need for a Multiyear Process for the Development of Performance Standards.** Future efforts to develop national consensus standards should not rely on highly constrained meetings and timetables. Instead, a national consensus process not unlike the 3-year effort to develop the NCTM Standards should be established.

6. **Provide for a Stable Basis for Comparison as Well as for Evolutionary Change.** Just as current curriculum frameworks are out of date, national content standards such as those being developed under the aegis of the National Academy of Sciences/National Research Council, the National Council of Teachers of English, and the like will need to be revised from time to time. Changes in assessment content and performance standards must necessarily follow. However, for national content standards to be feasible and useful, they must not change every 2 to 3 years. The Panel recommends a cycle of implementation, feedback, and revision that takes place over, perhaps, an 8- to 10-year period. Stability over some sustained period of time is important substantively, technically, and educationally. An 8- to 10-year cycle is long enough to establish meaningful trend lines, but short enough to accommodate the natural evolution that occurs in subject-matter fields.

Implications for the Design of NAEP

The recommendations made by the Panel regarding standard setting have important implications for other aspects of NAEP. The Panel has identified four issues that need to be examined further by the Governing Board and NCES to ensure that the quality of the National Assessment is not harmed by the effort to align NAEP with national content standards. The Panel urges that further consideration be given to each of the following:

1. To report progress toward the national education goals, it is essential that NAEP content domains and achievement levels match national content standards and national performance standards.

2. However, in addition to assessing forward-looking content standards, NAEP must also maintain itself as a comprehensive assessment of current practice. Therefore, explicit consideration will have to be given to the feasibility of blending "old" and "new" content in a single aggregate score.

3. Administration procedures may need to change to acknowledge the extra time demands of performance assessments. Some kind of branching technique may be required to make it possible to administer long and difficult problems to students who can do them without burdening students who cannot.

4. The issue of reporting by content subscales versus a single composite score should be re-evaluated for each subject area. If composite scores are formed, special attention should be given to their meaning for advanced and proficient work.

Final Observations of the Panel

The problems identified with the 1992 achievement levels are not technical quibbles. Large internal inconsistencies in judgments by the standard setters reflect major technical problems with the methodology. More important, external validity evidence suggests that use of the 1992 levels would lead to misinterpretation of NAEP results. If indeed the cutscores are set too high, then NAEP results for 1992 underreport the numbers of students attaining each level. This would mean, for example, that the public is being told that only a tiny percentage of American students are advanced, when in fact more students can do what is described by the advanced levels. At the same time, the content of the assessment does not reflect emerging content standards, particularly at the advanced level. Therefore, the public is presented with a very limited and low-level picture of what expectations should be for developing student thinking and mastery of challenging subject matter. On the one hand, more students meet the Governing Board's descriptions of advanced performance based on current content than was reported; on the other hand, this does not address how students

would do if measured against more demanding, emerging content standards. Reporting results based on flawed procedures risks undermining the credibility of the nation's primary indicator of educational achievement.

The Panel believes that a defensible procedure for setting performance standards is well within reach, due largely to the pioneering efforts of NAGB, its contractors, and the many evaluators of the 1990 and 1992 NAEP assessments. The Panel looks forward to the promulgation of rigorous and defensible achievement levels for NAEP, but cautions that it may take some time to establish them. To assist in reaching that objective, the Panel has recommended criteria and procedures for improving the interpretability and usefulness of NAEP reporting, for grounding NAEP in emerging national content standards, and for assuring continued credibility of NAEP as an essential indicator of achievement in American education.

1 National Context and Purpose for Setting Achievement Levels

It is unprecedented in the United States—where responsibility for education has traditionally and constitutionally been delegated to the states and local jurisdictions—for there to be widespread support for national standards for education. When the National Assessment of Educational Progress (NAEP) was created 25 years ago to monitor trends in education achievement, immense political pressure was expended to ensure that NAEP would not impose a national curriculum. Results were to be reported only by geographic regions and for the nation as a whole, not by state, and the project was *not* to be administered by a federal agency. But a dramatic change in attitudes toward national achievement standards occurred in the past 5 years, and pressure to establish such standards markedly increased. With a developing consensus for national education standards came the logical questions of who should set the standards and how progress against them should be measured. As one response, the National Assessment Governing Board (NAGB) decided to set achievement levels on NAEP beginning with the 1990 mathematics assessment. This report traces the history of the Governing Board's achievement-level activities and evaluates the setting of the achievement levels on the 1992 National Assessment.

Evolution of the National Education Standards Movement

Several factors have combined to make the notion of common national standards more popular and politically defensible. Beginning in 1983 with the appearance of *A Nation at Risk,* the rhetoric of education reform has created a close link between educational achievement and the economic competitiveness of the nation.[1] The fear of losing out in world markets in an age of a high-technology workforce has intensified the political pressure to reform elementary and secondary education. In addition, various analyses have made it increasingly clear that despite the absence of official policy, a de-facto national curriculum does exist and is driven by textbook marketing decisions and standardized tests.

The Education Summit and Creation of the National Education Goals Panel

Concerns about the educational preparation of the nation's youth prompted President Bush and the nation's governors to call an Education Summit in Charlottesville, Virginia, in September 1989. At this summit, President Bush and the nation's governors, including then Arkansas governor Bill Clinton, agreed on six broad goals

[1] U.S. Department of Education, *A Nation at Risk: The Imperative for Educational Reform* (Washington, D.C.: Author, 1983).

for education to be reached by the year 2000. Two of the goals pertained to student achievement with implications for the assessment of progress:

- *Goal 3:* By the year 2000, American students will leave grades 4, 8, and 12 having demonstrated competency in challenging subject matter including English, mathematics, science, history, and geography; and every school in America will ensure that all students learn to use their minds well, so they may be prepared for responsible citizenship, further learning, and productive employment in our modern economy.

- *Goal 4:* By the year 2000, U.S. students will be first in the world in science and mathematics achievement.[2]

Following the Education Summit, the National Education Goals Panel (NEGP) was created and charged with measuring progress toward the goals. The Goals Panel—six governors, four representatives from the Administration, and four members of Congress—launched six task forces and conducted hearings across the country. In attempting to implement these lofty goals, the Goals Panel was faced with a series of unprecedented questions. What is the challenging subject matter to be learned in the vast fields of English, mathematics, science, history, and geography? How is progress toward meeting these national goals to be measured? In particular, what kinds of assessments should be used, and what standards of performance should they contain? Should there be a single standard of competence or multiple standards?

Recommendations from the National Council on Education Standards and Testing

To pursue further the question of whether there should be national standards and examinations, Congress created the National Council on Education Standards and Testing (NCEST) in June 1991. NCEST delivered its report, *Raising Standards for American Education,* to Congress in January 1992.[3] NCEST endorsed the development of voluntary national content and performance standards and the development of school delivery standards. NCEST also envisioned a national state-of-the-art system of assessments to include both individual student and program assessments aligned with the national standards. Concluding that a mechanism was needed to coordinate the establishment of the standards and ensure quality control, NCEST recommended (1) that the Goals Panel be reconfigured to be more politically representative, and (2) that a new body be created to certify that the new standards, and the assessments based on them, would be world class.

In developing recommendations on education standards, NCEST drew a useful distinction between content standards and performance standards:

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

[3] National Council on Education Standards and Testing, *Raising Standards for American Education: A Report to Congress, the Secretary of Education, the National Education Goals Panel, and the American People* (Washington, D.C.: U.S. Government Printing Office, January 24, 1992, ISBN 0-16-036087-8).

> Content standards describe the knowledge, skills, and other understandings that schools should teach in order for students to attain high levels of competency in challenging subject matter.
>
> Student performance standards define various levels of competence in the challenging subject matter set out in the content standards.[4]

Content standards, which provide broad curricular goal statements, are exemplified by the *Curriculum and Evaluation Standards for School Mathematics* published by the National Council of Teachers of Mathematics (NCTM).[5] The NCTM Standards provide a broad outline of what should be taught in each level of schooling. For example, the 13 standards for grades K-4 include both topical standards for number sense, operations, computation, geometry, measurement, and recognizing patterns and relationships; and standards for the development of mathematical abilities such as problem solving, mathematical reasoning, communication, and making mathematical connections. Four illustrative standards for grades 5-8 are presented in figure 1.1. Although the NCTM Standards do not delineate specific instructional activities, they do set the direction for what should be taught. They imply a richer set of learning experiences and opportunities for deeper understanding of subject matter than are found in most traditional curricula. It is likely that the national content standards currently being developed in other subject areas will be similar to the NCTM Standards in level of generality.

Performance standards differ from content standards in their degree of specificity. Performance standards define just what students must do to demonstrate minimum, satisfactory, or superior levels of attainment. However, unlike the pass-fail performance standards created by minimum competency tests a decade or more ago, contemporary performance standards are expected to define high levels that most students will have to strive to attain. For example, performance standards could require successful completion of several different challenging assessment and performance tasks.

Content standards and performance standards are interdependent. Without performance standards to define what students must know and do to demonstrate their level of achievement, content standards may lack the forcefulness and specificity to motivate significant education reform. Performance standards that are predicated on partial or ill-formed content standards may encourage an undesirable narrowing of the curriculum to no more than the content tested, in the particular form in which it is assessed. Furthermore, performance standards based on inadequate content standards or low-level content run the risk of providing policymakers and practitioners with a false picture of what students actually can do. We return to this issue in chapter 5 of this report.

[4] Ibid., 13.

[5] National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (Reston, VA: Author, 1989).

Figure 1.1. ***Four examples of content standards from the NCTM standards, grades 5-8***

Standard 1: Mathematics as Problem Solving

In grades 5-8, the mathematics curriculum should include numerous and varied experiences with problem solving as a method of inquiry and application so that students can—

- use problem-solving approaches to investigate and understand mathematical content;
- formulate problems from situations within and outside mathematics;
- develop and apply a variety of strategies to solve problems, with emphasis on multistep and nonroutine problems;
- verify and interpret results with respect to the original problem situation;
- generalize solutions and strategies to new problem situations;
- acquire confidence in using mathematics meaningfully.

Standard 2: Mathematics as Communication

In grades 5-8, the study of mathematics should include opportunities to communicate so that students can—

- model situations using oral, written, concrete, pictorial, graphical, and algebraic methods;
- reflect on and clarify their own thinking about mathematical ideas and situations;
- develop common understandings of mathematical ideas, including the role of definitions;
- use the skills of reading, listening, and viewing to interpret and evaluate mathematical ideas;
- discuss mathematical ideas and make conjectures and convincing arguments;
- appreciate the value of mathematical notation and its role in the development of mathematical ideas.

SOURCE: National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (Reston, VA: Author, 1989).

Figure 1.1. ***Four examples of content standards from the NCTM standards, grades 5-8 (continued)***

Standard 3: Mathematics as Reasoning

In grades 5-8, reasoning shall permeate the mathematics curriculum so that students can—

- recognize and apply deductive and inductive reasoning;
- understand and apply reasoning processes, with special attention to spatial reasoning and reasoning with proportions and graphs;
- make and evaluate mathematical conjectures and arguments;
- validate their own thinking;
- appreciate the pervasive use and power of reasoning as a part of mathematics.

Standard 4: Mathematical Connections

In grades 5-8, the mathematics curriculum should include the investigation of mathematical connections so that students can—

- see mathematics as an integrated whole;
- explore problems and describe results using graphical, numerical, physical, algebraic, and verbal mathematical models or representations;
- use a mathematical idea to further their understanding of other mathematical ideas;
- apply mathematical thinking and modeling to solve problems that arise in other disciplines, such as art, music, psychology, science, and business;
- value the role of mathematics in our culture and society.

The Goals 2000 Initiative

The momentum for national standards continued in April 1993, when President Clinton transmitted the *Goals 2000: Educate America Act* to the Congress. Two of the major pillars of *Goals 2000* are to

- Codify into laws the six National Education Goals and a bipartisan National Education Goals Panel to report on progress toward achieving the goals.
- Establish a National Education Standards and Improvement Council to certify voluntary academic standards and assessments that are meaningful, challenging, and appropriate for all students.

The *Goals 2000* initiatives are crafted to provide national expectations or standards that are voluntary, having the power of persuasion to lead state and local activities. The national standards are to be broad, general frameworks leaving considerable latitude for local implementation. Further, the standards, developed by a consensus process involving national, state, and local groups, are to be national in character but not the product of a federal agency.

Even before *Goals 2000* was introduced, the setting of voluntary content standards in the nation was well underway. In addition to the mathematics standards developed by NCTM, standards for history, civics, geography, English, the arts, science, and foreign languages are being developed with funding from the U.S. Department of Education. This work is being conducted by various consortia of professional organizations, teachers, and researchers.

An Emphasis on "World-Class Standards"

The most important feature of the standards movement, which sets it apart from previous cycles of education reform, is its emphasis on high standards for all students. The phrase "world-class standards" is used to convey the idea that standards should be set high to be commensurate with standards in other nations and in contrast to minimum competency standards prevalent in the United States. The expectation is that all students can be taught to think and apply their knowledge to real-world problems. NCEST argued that challenging standards with a focus on substantive content, complex problem solving, and critical thinking would "raise the ceiling" for students who are currently above average and "lift the floor" for those experiencing the least success in school.[6]

The Need for New Forms of Assessments to Measure Challenging Curricula

The standards movement is closely tied to the need for new forms of assessment that embody the expectations of challenging curricula. In the 1970s and 1980s, education reform efforts placed a great deal of emphasis on basic skills, and outcomes were measured by standardized achievement tests. These were the goals emphasized by legislators and the public. A growing body of research evidence now suggests that such testing reinforced instructional efforts focused on rote learning and low-level skills. If it is understood that assessments serve to convey achievement goals and shape instructional practice as well as measure learning outcomes, then it is essential that assessment tasks accurately reflect curricular goals, both in form and substance. Better measures are needed to aim assessments at conceptual understanding, problem solving, and applying knowledge within and across subject areas. Thus implementation of the NCEST and *Goals 2000* vision will require assessments closely aligned with challenging content and performance standards.

[6] National Council on Education Standards and Testing, op. cit., 4.

Before a discussion of the role that NAEP is currently assuming in the assessment of national education standards, some additional background and history are needed.

NAEP as an Independent Monitor of the Status of Education

The National Assessment of Educational Progress, first administered in 1969, is conducted under the auspices of the National Center for Education Statistics (NCES). NAEP provides the best available information on the achievement of American students. First, its instruments are reasonably comprehensive measures of achievement in the core subject areas, including reading, writing, mathematics, science, history, and geography. Second, NAEP is administered to a probability sample of the nation's 4th, 8th, and 12th graders, thus assuring that the results are representative of what U.S. students know at any given time.

No other assessment or test provides results that are so comprehensive and representative. The widely known Scholastic Aptitude Test (SAT) and the American College Test (ACT) are not administered to national probability samples and thus cannot be used to evaluate national or state performance. For example, the states with the highest average SAT scores tend to be those with the smallest proportions of students taking the test because only their college-bound students take the SAT. It would be inappropriate to conclude that states with high average SAT scores are providing a higher quality education if, in fact, high scores only mean that average students do not take the test in those states. Standardized achievement tests such as the Comprehensive Tests of Basic Skills or Iowa Tests of Basic Skills are administered to national norm groups every few years, but due to the number of schools that refuse to participate in the norming, the samples are less trustworthy than those obtained for NAEP. Furthermore, as implied by their titles, these tests are not intended to measure advanced academic content.

NAEP Provides Invaluable Trend Data

NAEP items are administered to national samples under closely controlled conditions with precise checks on the comparability of scores from one year to the next. Thus NAEP provides invaluable trend data on the gains and losses in achievement by U.S. students. For example, NAEP trend data have shown that Hispanic and African-American students made appreciable gains over the past 20 years and that the achievement gap between majority and minority groups has narrowed. NAEP results have also documented a decline in the scientific knowledge of U.S. 17-year-olds. From the standpoint of the rigor of its data collection methods, NAEP is ideally suited for measuring progress toward education goals.

NAEP Must Maintain its Status as an Independent Monitor

In its first report, *Assessing Student Achievement in the States,* the Panel argued that "NAEP should be preserved as the nation's key independent indicator of how students are performing academically. NAEP has played and should play the same role for education that economic indicators have played in signaling changes in the status of the economy."[7] This recommendation is in keeping with recommendations from both the Goal 3 Resource Group and NCEST that NAEP serve the role of independent monitor in a system of multiple examinations. According to this view, a complete assessment system requires both an "independent indicator test ... designed to monitor the overall effectiveness of the education system, and individual student assessments designed to motivate student and teacher efforts to a high level of academic achievement."[8] The vision of a dual testing system with NAEP as the independent monitor has implications both for the relationship of NAEP to individual assessments and the conduct of NAEP itself.

NAEP Frameworks Must Be Comprehensive

As noted by NCEST and by the Panel, NAEP will have to be aligned with national standards as they are developed if NAEP is to help the nation measure progress toward achieving these standards. It is expected that NAEP frameworks will continue to be revised to make them more congruent with national content standards. However, as argued by the Panel in *Assessing Student Achievement in the States,* it is also important that a national indicator intended to measure change be as comprehensive as possible. That is, NAEP frameworks should be broad enough to reflect goals for what students should know and be able to do as well as how they are presently being educated. NAEP content frameworks and test items should not shift wildly in response to fads within subject-matter disciplines, or NAEP will be unable to measure changes in learning. For example, in the 1980s, NAEP provided clear evidence that students gained in rote skills (emphasized in previous rounds of reform) while losing on higher-order thinking skills. If earlier reformers had insisted that NAEP measure only basic-level skills, the data would never have been collected to make the trade-off in learning goals apparent.

NAEP's role as an independent monitor and its relation to the development of national standards are important when considering each new technical and policy question about the assessment, including the evaluation of NAEP's achievement levels. The NAEP achievement levels, along with their verbal descriptions and exemplar items, must also be judged in terms of their congruence with national content standards and their adequacy for communicating results to the public. If the purpose for gathering sound data is to inform public policy and contribute to education reform, then both the technical adequacy of education indicators and their accuracy and clarity of interpretation must be considered.

[7] The National Academy of Education, *Assessing Student Achievement in the States* (Stanford, CA: Author, 1992), 67.

[8] National Education Goals Panel, *Measuring Progress Toward the National Education Goals: Potential Indicators and Measurement Strategies* (Washington, D.C.: Author, March 25, 1991), 48.

NAGB's Decision to Set Achievement Levels

Congressional Authorization

As part of the NAEP reauthorization in 1988, the National Assessment Governing Board was created to set consistent policy direction for the assessment. Prior to 1988, the law provided that the NAEP contractor appoint a separate Assessment Policy Committee to establish policy for each assessment. The legislation assigned NAGB the following responsibilities:

- Selecting subject areas to be assessed;
- Identifying appropriate achievement goals for each age and grade in each subject area to be tested under the National Assessment;
- 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; and
- Taking appropriate actions needed to improve the form and use of the National Assessment.[9]

NAGB's Interpretation of the Law

The charge from Congress to NAGB to identify appropriate achievement goals in each subject area is cited by NAGB as the authority for its decision to set achievement levels for NAEP assessments. NAGB might have responded to the authorizing legislation in different ways. The charge could have been interpreted as a mandate to develop broad curricular goals in the spirit of the NCTM Standards. This would have meant using the consensus approach to determine the content frameworks as NAEP has always done, without further specifying performance standards. Or NAGB might have assumed that the Goals Panel would set performance standards and that NAEP would be called on to measure progress against these goals. Instead NAGB took the charge to mean they should first set performance standards and then measure progress against them using the NAEP assessment scores.

[9] Public Law 100-297, Part C, Section 3403 (6) (A): 1988.

The Governing Board's decision to set achievement levels is best understood in the context of the push for national education standards that followed from the Education Summit. In a NAGB staff paper dated December 8, 1989, Roy Truby, the Executive Director, reviewed the 1988 legislation creating the Governing Board and its responsibility for "identifying appropriate achievement goals" alongside the call at the Education Summit for national education goals and an annual report card to monitor achievement toward these goals. He then argued that because NAEP is the only reliable national measure of achievement in the cognitive domain, the statements from the Summit imply "that NAEP should be involved in setting substantive achievement standards."[10] He concluded that NAGB should "undertake to develop grade-level achievement goals that are 'under the umbrella' of broad national goals."[11]

The Use of NAEP as a Lever for Education Reform

One of NAGB's intentions in launching the effort to set achievement levels was to make NAEP results more understandable, more potent with the public and with policymakers, and hence more useful in advancing education reform. An early NAGB staff paper suggested that reporting results in terms of the proportions of students who attain substantive standards "may serve as a lever for school improvement."[12] According to then NAGB Chairman Chester E. Finn, Jr., "NAEP has long had the potential not only to be descriptive but to say how good is good enough....In the spirit of Charlottesville the Board has now started moving to establish benchmarks for learning, to use with its tests."[13] Setting achievement levels on NAEP was seen as a way to participate in the national agenda for setting goals and to enhance the usefulness of NAEP for measuring progress toward the goals.

The Goals Panel's Emphasis on Achievement Levels

As previously described, the Goals Panel, established in 1990 by President Bush and the National Governors' Association, was given the charge of reporting, at least annually, progress toward achieving the national education goals. The Goals Panel immediately began plans to use NAEP results, nationally and state by state, in its reports. Furthermore, the Goals Panel determined that its own reports of NAEP results should use *and emphasize* achievement levels in setting expectations for progress needed to reach the national goals by the year 2000. These determinations were important in reinforcing NAGB's decision to develop achievement levels for reporting the results of the 1990 mathematics assessment.

[10] Roy Truby, "Staff Paper on Setting Goals for the National Assessment" (Washington, D.C.: National Assessment Governing Board, December 8, 1989), 4.

[11] Ibid., 5.

[12] Roy Truby, "The Future of NAEP as the Nation's Report Card" (Washington, D.C.: National Assessment Governing Board, November 29, 1989), 10.

[13] C.E. Finn, Jr., news release (Washington, D.C.: National Assessment Governing Board, December 12, 1989), 1.

Opposition to the Use of Achievement Levels

NAGB held several public hearings on their proposal to report NAEP results using achievement levels. Although much of the testimony was favorable, not all groups were in favor of establishing achievement levels on NAEP. One of the groups to raise questions was the Council of Chief State School Officers (CCSSO). At a January 25, 1990, forum, CCSSO stated their belief that although NAEP could be used to measure progress against some of the national goals, a distinction had to be made "between the use of the NAEP results as a national goal in itself and the use of NAEP to measure progress toward a national goal." CCSSO went on to say, "The NAGB Board has been assigned no statutory responsibility for establishing national goals for education."[14,15] An *Education Week* article cited warnings from testing experts about technical obstacles in the setting of achievement levels, the difficulty of making predictions about prerequisites for college-level work without empirical studies, and concerns about whether NAGB was the appropriate agency to determine national standards.[16] In public hearings, objections to an implicit national curriculum were raised by representatives of the National Conference of State Legislatures and the National Association of Secondary School Principals. In contradictory testimony, objections were raised to a single standard because it would encourage effort only at the low end of the distribution, and to three standards because they would reinforce current tracking and the practice of holding different expectations for different students.[17]

Specification of Three Achievement Levels

In the presence of these objections and concerns, NAGB went forward with a plan to specify achievement levels for the 1990 mathematics assessment at grades 4, 8, and 12. They first adopted generic descriptions of basic, proficient, and advanced levels, presented in figure 1.2. These generic descriptors served as the basis for more specific, substantive descriptions of the basic, proficient, and advanced levels in each subject area and at each of the three grade levels. NAEP scale cutpoints representing marginally basic, marginally proficient, and marginally advanced performance at each grade level were then determined. The cutpoints defined four score groups for each grade, the lowest being "below basic." A further description of the levels and the process by which they were set is provided in chapter 2.

[14] Council of Chief State School Officers, "Comments of the Council of Chief State School Officers on 'Setting Goals for the National Assessment'" (Presentation at a National Assessment Governing Board Forum, January 25, 1990).

[15] Some significant changes were made in the achievement-level-setting process between the 1990 and 1992 assessments. As a result, the use of achievement levels as the primary means for reporting NAEP results for the 1992 administrations of mathematics and reading was endorsed by state testing directors through the Education Information Advisory Committee (EIAC) Assessment Subcommittee convened by the Council of Chief State School Officers. In May 1992, EIAC recommended that (1) both achievement-level data and anchor-level data be produced for the 1992 NAEP Trial State Assessment of mathematics; and (2) the results be presented in the report primarily in terms of achievement levels as done in the draft prototype reports. This resolution was endorsed by the Council of Chief State School Officers as well.

[16] R. Rothman, "NAEP to Create Three Standards for Performance," *Education Week* 9 (35) (May 23, 1990): 1.

[17] R. Rothman, "NAEP Plan to Set Performance Goals Questioned," *Education Week* 9 (19) (January 31, 1990): 19.

Figure 1.2. ***Achievement-level definitions adopted by NAGB on May 11, 1990***

Basic

Denotes partial mastery of the knowledge and skills that are fundamental for proficient work at each grade—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

Represents solid academic performance for each grade tested—4, 8, and 12—and reflects a consensus that students reaching such a level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. At grade 12, the proficient level will encompass a body of subject-matter knowledge and analytical skills, of cultural literacy and insight, that all high school graduates should have for democratic citizenship, responsible adulthood, and productive work.

Advanced

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.

SOURCE: National Assessment Governing Board, *Achievement Level Options for the NAEP Mathematics Assessment: 1990 Trial* (Washington, D.C.: Author, May 10, 1991).

It is important to note that the development of the achievement levels for the 1990 mathematics assessment occurred *after* the NAEP test items had been developed and administered. The assessment framework and items had not been designed to yield results by achievement level; rather, the levels were created in retrospect.

The 1990 NAEP results were originally reported in June 1991 in their usual manner, without reference to the achievement levels. In September 1991, the Governing Board issued a separate report in which the three levels—basic, proficient, and advanced—were used. At the same time, the Goals Panel included the same results in their 1990 "report card," but chose to present them by only two levels—competent and not competent—thereby reflecting a disagreement about the formulation of the Governing Board's levels. Within a few months, the public had seen the same assessment results reported in three different forms.

The Need to Evaluate the Achievement Levels

The decision to use achievement levels as the primary lens for reporting NAEP results can be seen as another in a series of steps to make NAEP more interpretable and useful. For example, the decision to introduce a new scaling procedure in 1984 was made to permit the reporting of a single score in each subject area.[18] A single score was intended to make it easier for policymakers and the public to understand assessment results. The move to state-level reporting, which had been recommended in the Alexander-James report, *The Nation's Report Card,* was also made for policy purposes, given that states are responsible for most education policy decisions.[19] Other proposals, such as reporting NAEP results at the district level, are again intended to make NAEP results more visible and useful.

Each of these changes is controversial because of potential threats to the quality and integrity of NAEP data. These concerns were the reason, in fact, that the Trial State Assessment (TSA) was created as a "trial" with a formal evaluation planned from the outset. In its first report of the TSA evaluation, however, this Panel found no serious threats to the quality of test administration in the state-by-state NAEP. However, each important shift in the design of NAEP, including reporting by achievement levels, must be evaluated to determine both intended and unintended consequences.

Given the national context, purposes for establishing achievement levels, and NAEP's role as an independent monitor, the Panel's evaluation of the NAEP achievement levels addresses three critical issues:

- Does reporting scores in terms of achievement levels accomplish its intended purpose of better communicating NAEP results to the public?

[18] The procedure used is called Item Response Theory (IRT). A description of this theory and its application to NAEP can be found in S. Messick, A. Beaton, and F. Lord, *National Assessment of Educational Progress Reconsidered: A New Design for a New Era,* NAEP Report No. 83-1 (Princeton, NJ: Educational Testing Service, 1983).

[19] Lamar Alexander and H. Thomas James, *The Nation's Report Card: Improving the Assessment of Student Achievement* (Washington, D.C.: National Academy of Education, 1987).

- Is the change in reporting consistent with overall national efforts to establish education standards?
- Can NAGB's intended policy goals be accomplished without negatively affecting the accuracy and integrity of NAEP data?

The National Academy of Education Panel on the Evaluation of the Trial State Assessment

The National Academy of Education (NAE) Panel was created in 1990 following a mandate from Congress for an independent evaluation of the extension of NAEP to the state level. In its 1991 evaluation report, the Panel addressed a series of issues pertinent to the technical accuracy and policy relevance of the 1990 mathematics Trial State Assessment. Due to the salience of the standards issue and controversy surrounding earlier evaluations of the achievement levels, in 1992 NCES asked the Panel to expand its work to provide an extensive evaluation of the achievement levels and the issues surrounding their use. The chair of NAGB's achievement-level committee welcomed the Panel's observation and evaluation of the level-setting process, and draft legislation for the reauthorization of NAEP stipulated that the achievement levels be evaluated by the Panel.[20] In response, this report of the NAE Panel on the Evaluation of the NAEP Trial State Assessment focuses specifically on the validity and usefulness of the NAEP achievement levels. Empirical findings from commissioned studies and the Panel's own analyses and evaluation are presented in subsequent chapters of this report.

Some points treated in the Panel's validity studies are specialized and technical. These technical points are critical to understanding the adequacy and appropriateness of interpretations and uses of the NAGB achievement levels. However, the fundamental issues addressed in this report are not technical ones. Instead, we consider what it means to set standards for a nation. If statements of curricular expectations and samples of student work are intended to shape public discourse and determine education policy, how will we determine whether content and performance standards are appropriate and adequate for these profound purposes? Do benchmarks set on the NAEP test represent the important aspirations held for the nation's schools and report accurately the proportions of students attaining those goals? Technical considerations should be evaluated in light of these broader questions.

Two themes are critically important in defining the policy context for this evaluation. First, the establishment of achievement levels for NAEP must be done in relation to the national effort to set education standards. Second, the role of NAEP as an independent monitor must be maintained. NAEP, as the nation's indicator of educational attainment and progress, must continue to assess present conditions and also assess achievement on new content standards that reflect the country's aspirations for standards-based curriculum, instruction, and assessment.

[20] On February 4, 1992, Michael Glode, then Chairman of NAGB's achievement-level committee, sent a letter to Robert Glaser and Robert Linn, co-Chairmen of the NAE Panel on the Evaluation of the Trial State Assessment, welcoming the Panel's evaluation of NAGB's setting of the achievement levels for the 1992 assessment.

Finally, the charge to the Panel to evaluate the 1992 NAEP achievement levels was, in part, a consequence of earlier critical evaluations of the setting of the achievement levels. It is important to state that the Panel's evaluation of the 1992 standard-setting effort is based on new evidence entirely independent of results of earlier evaluations. Although some of the technical procedures the Panel has adopted and the manner in which some specific issues are framed have capitalized on the procedures used in the evaluations of the 1990 achievement levels, our focus has been solely on the setting of the 1992 levels. To the extent that the Panel's findings corroborate the conclusions of past evaluations, the conclusions of each of the evaluations are strengthened.

An Overview of this Report

The evaluation's overarching research question is whether NAEP reports that make use of the 1992 NAEP achievement levels will lead to valid interpretations of NAEP results. To evaluate the achievement levels and to provide information that could be used to improve the achievement-level-setting process in the future, the Panel commissioned a series of studies. The Panel's commissioned studies were designed to evaluate standard setting in the areas of reading, mathematics, and writing—the three areas in which achievement levels were set in 1992.[21] One set of studies focused on internal analyses of the standard-setting process and are reported in chapter 3. A second set of external comparison studies are reported in chapter 4. Before these empirical findings are presented, additional information specific to standard-setting procedures is provided in chapter 2.

Chapter 2 addresses score reporting and NAGB's achievement levels. The traditional method for reporting NAEP results is explained, with particular attention to the techniques used to give meaning to the score scale. Then research literature on setting cutscores is reviewed, leading up to the rationale for NAGB's decision to use a modified Angoff procedure. A brief summary of previous evaluations of NAGB's achievement levels is given, followed by an overview of the specific procedures used in 1992 to set the levels in reading and mathematics. The achievement-level descriptions and cutpoints on the NAEP scale are presented for both reading and mathematics, with an example of reporting by achievement levels from the 1992 mathematics assessment.

In chapter 3 the findings on the internal validity of the achievement levels from the Panel's studies of the standard-setting process itself are reported. These studies involve reanalysis of data collected from the judges by NAGB's contractor during different rounds of the standard-setting process, observational data from the standard-setting meetings, observations of the followup meetings to refine descriptions of the levels, and surveys sent to participants. In addition, the Panel conducted independent experimental studies to examine the effects of certain features of the standard-setting procedures, such as focusing on item-level judgments and setting three cutpoints in order. Taken together, these studies inform the Panel's evaluation of whether the process was internally consistent, coherent, and free from artifactual biases, and whether it made sense to the participants and led to consensual decisions.

[21] However, ACT, the NAGB contractor, encountered a series of problems which delayed the release of the writing data. As a result, the Panel decided not to include an evaluation of the writing achievement levels in this report. Because of the problems encountered, the Governing Board decided on July 8, 1993, not to report the 1992 writing data using achievement levels.

In chapter 4 the validity and reasonableness of the final levels identified in reading and mathematics are addressed. Although no external criteria can serve as the ultimate authority in judging what constitutes advanced, proficient, or basic performance, external comparisons can provide perspective on the reasonableness or unreasonableness of the final standards, and these are described. The Panel also conducted extensive field-based studies to gather collateral performance data to evaluate whether classifications based on classroom performance correspond to classifications based on NAEP achievement levels. National results from Advanced Placement examinations provided another look at the reasonableness of "advanced" classifications at grade 12. Also, for eighth-grade mathematics, it was possible to compare U.S. results using NAGB's standards to results in other countries. Finally, panels of subject-matter experts were asked to evaluate the meaningfulness of the achievement levels in light of the current thinking in each field, especially those curricular expectations and aspirations likely to be reflected in national content standards. For example, the mathematics experts considered the appropriateness of the level descriptions and exemplar items and their effectiveness in communicating NAEP results in light of the NCTM Standards.

In the fifth and final chapter, the Panel presents its findings and recommendations on the development of NAEP achievement levels. The desirability of having achievement levels is considered, and alternate ways for reporting are discussed. Issues not addressed in the empirical work, such as the advisability of separate scales for each grade and conceptions of what it means to have advanced knowledge or expertise, are identified. The chapter concludes with the Panel's recommendations on the essential elements for further development and use of performance standards and achievement levels. These recommendations include long term linking of national content standards and performance standards.

2 NAEP Score Reporting and Standard Setting

NAGB's achievement levels are the first highly visible examples of national education performance standards. As such, their potential impact reaches beyond their direct influence on NAEP reporting and embraces the likelihood that these standards, or the process used to set them, could become models for future efforts. It is, therefore, both understandable and appropriate that NAGB's standard-setting project has been evaluated more than any other aspect of NAEP.

This chapter provides background for the Panel's evaluation of the 1992 achievement levels. Traditional methods used to report NAEP results are described first. This is followed by summaries of the tests and measurements literature on standard setting and of NAGB's stated rationale for using a modified Angoff procedure. In the fourth section, a summary of issues identified in previous evaluation studies is provided. An overview of the specific procedures used to set the levels for 1992 and to create the level descriptions in reading and mathematics is given next. And finally, the achievement-level descriptions and cutpoints are presented, along with a discussion of some of the issues that arose in the reporting of the 1992 achievement levels.

Traditional Reporting of NAEP Results

The method of reporting results on NAEP has evolved over time. Because students do not all take the same test and because test items change from year to year, it is not informative simply to report NAEP results as the number right or percentage of right answers. In the early years, NAEP reported results item by item, publishing the text of the item along with information on the proportion of students who answered correctly. This method of reporting made it immediately apparent what students could do on selected NAEP items because the specific test question was shown. However, it also resulted in a proliferation of results with no overall summary provided. Item-by-item reporting was replaced by average p-values for sets of items judged to be similar with respect to the knowledge or skills measured.[1] These item groups were replaced, in turn, by a sophisticated scaling procedure that made it possible to summarize performance for a whole subject area with a single composite score. Scaling also enabled comparisons across grades and across years without having to administer precisely the same sets of items.

The current NAEP scaling procedure, in use since 1984, produces a standard score scale ranging theoretically from 0 to 500. The scale is initially established to have a mean of 250 and standard deviation of 50 (for grades 4, 8, and 12 combined) in the base year. After the base year, the scale is held constant so that changes from year to year can be observed. A new base year is set each time the content framework for a subject area is changed. Thus 1990 was a new base year for mathematics, and 1992 was a new base year for reading.

[1] p-value is a measurement term referring to the proportion of examinees who answer an item correctly.

To enhance the interpretability of NAEP scores, Educational Testing Service (ETS), the NAEP contractor, provides descriptions of the kinds of things that students know and can do at each major score interval. These descriptions traditionally have been provided for so-called "anchor points" corresponding to standard deviation intervals above and below the mean: 200, 250, 300, 350.

The anchor-point descriptions are generated empirically after the NAEP administration by carefully examining anchor items—items answered correctly by a substantial proportion (65 percent) of the students whose overall performance places them at a given anchor point. To ensure accurate distinctions between anchor points, the methodology imposes two additional statistical requirements: (1) that an anchor item not be answered correctly by a majority of students at the next-lower anchor point and (2) that the change in percent correct for anchor items be at least 30 percentage points between the next-lower anchor point and the intended anchor point. Moving up the NAEP scale, anchor items thus exemplify what students at a given point can do that students at the point below cannot do.

Once anchor items are identified, a group of content specialists reviews them and writes general descriptions of the content and processes they entail, summarizing the capabilities they represent. These anchor-point descriptions are published along with illustrative anchor items to give meaning to the scale and improve its interpretability. Figure 2.1 provides an example of the 1992 anchor-point descriptions in mathematics, shown in relation to the NAEP scale.

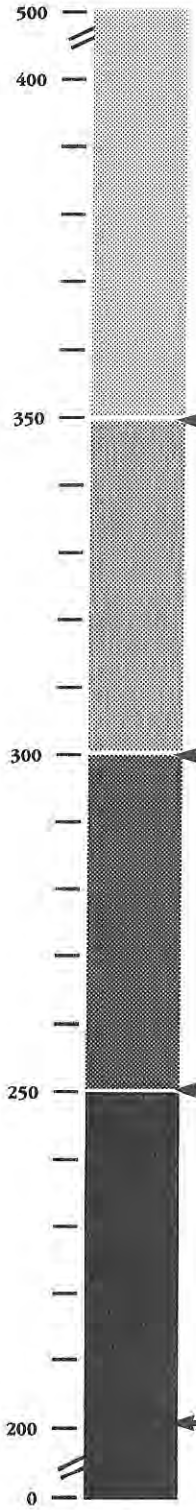
In recent years it has been customary to report NAEP results both as an average score on the NAEP scale and in terms of the percentage of students in each grade scoring above each of the anchor points. Table 2.1 shows these data for the 1990 and 1992 mathematics assessments. As would be expected when a single scale is used across grades, the data show that proficiency increases with grade level. In 1992 only 17 percent of 4th graders were at or above a score of 250, whereas 68 percent of 8th graders and 91 percent of 12th graders were at or above this point. By contrast, the achievement levels, being developed separately for each grade, produce similar percentages of students at each level within grade. For example, in 1992, as can be seen in table 2.2, 2 percent of 4th graders were advanced (by 4th-grade standards), 4 percent of 8th graders were advanced, and 2 percent of 12th graders were advanced.

Although the anchor-point descriptions are useful for characterizing what students at each anchor point can do, they do not, of themselves, provide a goal or standard against which student performance can be evaluated. In the past, various efforts have been made to add an evaluative component to the anchor points, but none have survived. For example, in 1984, labels such as advanced (350), adept (300), intermediate (250), basic (200), and rudimentary (150) were used to characterize performance at the respective anchor points. Users did not find this labeling helpful, however, because it was applied to a cross-grade scale. For example, the anchor-point labels did not provide a meaningful way to describe superior performance at the lower grade levels.

More recently, subject-matter committees added information about the grade equivalencies of the anchor points to the 1990 mathematics report. For example, anchor point 250 was said to be representative of material generally covered by the fifth grade.[2]

[2] National Center for Education Statistics, I.V.S. Mullis et. al., *The STATE of Mathematics Achievement* (Washington, D.C.: Author, June, 1991), 6-7.

Figure 2.1. ***Description of mathematics proficiency for four anchor levels on the NAEP scale***



350–Reasoning and problem solving involving geometric relationships, algebra, and functions.

Students at this level can reason and estimate with percents. They can recognize scientific notation and find the decimal equivalent. They can apply their knowledge of area and perimeter of simple geometric figures to solve problems. They can find the circumferences of circles and the surface areas of solid figures. They can solve for the length of missing segments in more complex similarity situations. Students can apply the Pythagorean Theorem to find the hypotenuse of a right triangle. They are beginning to use rectangular coordinates in problem-solving situations and can apply geometric properties and relationships in solving problems.

Students can compute means from frequency tables and create a sample space to determine probabilities, and read the graph of a step-function. Students can use exponents and evaluate expressions given in functional notation. In number theory, they have an understanding of even and odd numbers and their properties. They can identify an equation describing a linear relation provided in a table, and solve literal equations and systems of two linear equations. They have some knowledge of trigonometric relations. These students can represent and interpret complex patterns and data using numbers, expressions, and graphs. Given the graph of a function they can identify its zeros and the effect on the graph of taking the absolute value of the function.

300–Reasoning and problem solving involving fractions, decimals, percents, elementary concepts in geometry, statistics, and algebra.

Students at this level can use various strategies and explain their reasoning in a variety of problem-solving situations. They are able to solve problems involving not only whole numbers but with decimals and fractions. They can represent and find equivalent fractions, and use these concepts in solving routine problems. They can find a percent of a number and use this skill in simple problems. Multiplication and division of whole numbers have developed to the extent that students can use all four operations in multi-step problems.

Students can read and use instruments in more complex situations. They can find areas of rectangles, recognize relationships among common units of measure, and solve routine problems involving similar triangles and scale drawings. They have knowledge of definitions and properties of simple geometric figures in the plane. Their spatial sense includes the ability to visualize a cube in either three-space or its flattened form in a plane.

Students can calculate averages, select and interpret data from a variety of graphs, list the possible arrangements in a sample space, find the probability of a simple event, and have a beginning understanding of sample bias. They can use knowledge of relative frequencies in simple simulation situations. Students show the ability to evaluate simple expressions and solve linear equations. Students can graph points on coordinate axes, locate the missing coordinates for a corner of a square, and identify which ordered pairs satisfy a given linear equation.

250–Multiplication and division, simple measurement, and two-step problem solving.

When presented with a problem situation, students at this level have some understanding of the problem, can identify extraneous information, and have some knowledge of when to use computational estimation. They have an understanding of addition, subtraction, multiplication, and division with whole numbers. They can solve simple two-step problems involving whole numbers. They are able to round whole numbers and solve simple word problems involving place value, estimation, and multiples.

Students can use a ruler to measure length in centimeters and have some understanding of area and perimeter. They can solve simple problems using readings from instruments. They demonstrate a knowledge of properties of triangles, squares, rectangles, circles, and cubes. They can solve problems that require visualizing, drawing or manipulating simple geometric shapes. They are able to complete bar graphs and pictographs, as well as use information from graphs or tables to solve simple problems. They can recognize simple number patterns, are beginning to deal informally with the idea of a variable, and have some knowledge of simple probability.

200–Addition and subtraction, and simple problem solving with whole numbers.

Students at this level can identify solutions to one-step word problems, involving addition or subtraction. They can add and subtract whole numbers in most situations, and when a calculator is available, they can multiply and divide. They are able to select the largest whole number from a set of numbers in the thousands, and can match the verbal and symbolic names for numbers.

Students demonstrate familiarity with length and weight, by selecting appropriate instruments and units to measure these attributes. They are able to recognize some basic properties of two-dimensional geometric figures as well as the names of standard examples of these figures. They can recognize simple patterns.

Table 2.1. ***Example of reporting by anchor points for the 1990 and 1992 mathematics assessments: National overall average mathematics proficiency and anchor points, grades 4, 8, and 12***

		Assessment Years	Grade 4	Grade 8	Grade 12
Average Proficiency		1992 1990	218(0.7)> 213(0.9)	268(0.9)> 263(1.3)	299(0.9)> 294(1.1)
Anchor Point	Description	Percentage of Students At or Above			
200	Addition and Subtraction, and Simple Problem Solving with Whole Numbers	1992 1990	72(0.9)> 67(1.4)	97(0.4) 95(0.7)	100(0.1) 100(0.2)
250	Multiplication and Division, Simple Measurement, and Two-Step Problem Solving	1992 1990	17(0.8)> 12(1.1)	68(1.0) 65(1.4)	91(0.5)> 88(0.9)
300	Reasoning and Problem Solving Involving Fractions, Decimals, Percents, and Elementary Concepts in Geometry, Statistics, and Algebra	1992 1990	0(0.1) 0(0.1)	20(0.9)> 15(1.0)	50(1.2)> 45(1.4)
350	Reasoning and Problem Solving Involving Geometric Relationships, Algebra, and Functions	1992 1990	0(0.0) 0(0.0)	1(0.2) 0(0.2)	6(0.5) 5(0.8)

> The value for 1992 was significantly higher than the value for 1990 at about the 95 percent confidence level. < The value for 1992 was significantly lower than the value for 1990 at about the 95 percent confidence level. The standard errors of the estimated percentages and proficiencies appear in parentheses. It can be said with 95 percent confidence that for each population of interest, the value for the whole population is within plus or minus two standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference. When the proportion of students is either 0 percent or 100 percent, the standard error is inestimable. However, percentages 99.5 percent and greater were rounded to 100 percent and percentages 0.5 percent or less were rounded to 0 percent.

SOURCE: I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *NAEP 1992 Mathematics Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, 1993), 235.

Table 2.2. ***Example of reporting by achievement levels for the 1990 and 1992 mathematics assessments: National overall mathematics proficiency and achievement levels, grades 4, 8, and 12***

Grades	Assessment Years	Average Proficiency	Percentage of Students At or Above: Advanced	Proficient	Basic	Percentage Below Basic
4	1992	218(0.7)>	2(0.3)	18(1.0)>	61(1.0)>	39(1.0)<
	1990	213(0.9)	1(0.4)	13(1.1)	54(1.4)	46(1.4)
8	1992	268(0.9)>	4(0.4)	25(1.0)>	63(1.1)>	37(1.1)<
	1990	263(1.3)	2(0.4)	20(1.1)	58(1.4)	42(1.4)
12	1992	299(0.9)>	2(0.3)	16(0.9)	64(1.2)>	36(1.2)<
	1990	294(1.1)	2(0.3)	13(1.0)	59(1.5)	41(1.5)

> The value for 1992 was significantly higher than the value for 1990 at about the 95 percent confidence level. < The value for 1992 was significantly lower than the value for 1990 at about the 95 percent confidence level. The standard errors of the estimated percentages and proficiencies appear in parentheses. It can be said with 95 percent confidence that for each population of interest, the value for the whole population is within plus or minus two standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference.

SOURCE: I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *NAEP 1992 Mathematics Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, 1993), 64.

The grade equivalencies were discontinued after concerns were raised about the appropriateness of such judgments, given that they were neither derived by a consensus process nor systematically reflective of the curriculum of the participating states.

It was partially in light of these concerns and criticisms of the anchor points that NAGB decided to add another dimension to NAEP reporting, namely, the standards-based achievement levels.

Research Literature on Setting Cutscores

In this section we review the technical history of standard setting in order to provide a context for NAGB's decision to use the modified Angoff procedure for setting standards on NAEP. The tests and measurement research literature on standard setting has generally used the term "standard" to refer to a passing score (cutscore) on a test,

typically a minimum competency test such as those inaugurated in the 1970s, or a test to establish professional certification or licensure. Cutscores answer the question "How good is good enough?" and are more or less analogous to performance standards as defined by NCEST and described in chapter 1. (The literature on standard setting does not generally address the related topic of curricular and content standards development as it is occurring in the wider effort to create national education standards.)

Literally dozens of standard-setting methods are described in the technical literature. Most can be reduced to a few basic approaches, which we consider here under the global categories of judgmental and empirical methods.[3]

Judgmental Methods

Jaeger referred to judgmental methods as test-centered methods because they involve judges setting cutscores (also called cutpoints) after closely studying the items on a test.[4] The Angoff procedure is the most widely used standard-setting method and the most straightforward of the judgmental procedures.[5] The gist of the Angoff method in the context of minimum competency testing is roughly as follows. First, expert or lay judges must conceptualize what they think is minimally competent performance (or a just-barely-passing knowledge level) on the content domain measured by the test. Then, holding in mind an image of an individual who just surpasses this level of minimal competency, judges must review each item on the test and estimate the probability that that minimally competent individual would answer the item correctly. Alternatively, judges may be asked to imagine a group of, say, 100 minimally competent individuals and estimate what proportion of them would get the item correct.

A recommended passing score, in terms of the number of items answered correctly on the total test, can be derived for any given judge by summing across the probabilities (p-values) that that judge estimated for each of the individual items on the test. Usually, individual judges' scores are then averaged to come up with an overall recommended standard or cutpoint. Implementation of this basic approach varies tremendously from study to study, depending upon the instructions given to judges and the information provided. It is common for the Angoff procedure to involve several rounds of item ratings, with judges being given new information at each stage, such as the ratings of other judges and student item-performance data.

[3] In addition, standard-setting methods can be classified as state or continuum models. State models are relevant when a true and absolute distinction exists between mastery and nonmastery states. Such dichotomies rarely occur in practice, except for single, unitary skills such as being able to add two-digit numbers. Determining achievement levels for complex and heterogeneous subject matters such as those assessed by NAEP clearly involves segmenting continuous performance scales. Therefore, state standard-setting models are omitted from the discussion.

[4] R.M. Jaeger, "Certification of Student Competence," *Educational Measurement,* 3rd ed., Ed. R.L. Linn (New York: American Council on Education and Macmillan, 1989).

[5] More information is given in a later section of this chapter about the modified Angoff method used by NAGB.

Empirical Methods

The dichotomy between judgmental and empirical methods is artificial in the sense that both methods involve judgments and both rely on empirical data. Empirical methods (referred to by Jaeger as examinee-centered methods), however, require judges to make judgments about examinee performance outside of the test context.[6] The contrasting-groups method is the best known of the empirical methods. It can also be used to help evaluate the validity of judgmentally set standards.

To implement the contrasting-groups method, judges are asked to evaluate examinees based on their real-world performance. For example, teachers might be asked to identify all the students in their classes who are at least minimally competent writers (according to an agreed-upon definition) as well as those who are not competent. The test for which the cutpoint is being developed would then be administered to all of the students, and a cutpoint would be established based on comparisons of the actual test performances of the competent and noncompetent students. Various statistical techniques can be used to set an "optimal" cutpoint, that is, one which comes the closest to making the same classifications on the test as were made based on real-life performance. Again there are many different variations in the implementation of this method. For example, to improve classification accuracy, judges can be encouraged to leave out individuals who are on the borderline between competence and noncompetence.

Normative Considerations and Decision-Theory Adjustment Methods

Normative considerations may seem antithetical to the notion of absolute standards. After all, standards are intended to reflect expectations based on substantive choices, not the norm or average performance. Nevertheless, there are several ways in which normative information can (and does) inform the standard-setting process without the norms becoming the standards. First, absolute standards are implicitly affected by normative understandings of what is possible. For example, if standards were determined only by what was desirable, then why not set eighth-grade expectations for fourth graders? In fact, what is possible (which is informed by normative data) sets implicit boundaries for reasonable expectations. This same logic explains why judges implementing the Angoff method are sometimes given data on item p-values, (i.e., on the proportion of examinees who get an item right).

It is also relevant to ask standard setters directly what percentage of examinees should fail the test. Sometimes this question has neither substantive nor normative implications, as when passing rates might be set to control the number of new candidates entering the profession. Judges may, however, bring relevant normative information to bear that is equivalent to external validity evidence. For example, judges being asked to set a passing score on a veterinary licensure examination might be asked to think about what percentage of students in the state veterinary school appear to be competent to practice. Assuming that nearly every student in the veterinary school takes the test, a valid cutpoint on an otherwise valid test would then

[6] Jaeger, op. cit.

be one that passes a similar proportion of examinees. For another example, high school graduation standards might be informed by data on the percentage of students required to take remedial courses in college.

Finally, a great number of the methods proposed in the standard-setting literature are actually adjustment models, intended to adjust standards (cutscores) after they are set initially by judgmental or empirical methods. No test score perfectly predicts an individual's performance, and no individual is likely to get exactly the same score every time he or she takes a test. Therefore, some error is always built into the use of test-based standards to certify competence. Adjustment models allow one to manipulate the kinds of errors that will occur.

For example, in the case of certifying surgeons or airline pilots, the risk to the public of falsely certifying someone who is not really competent makes this the more serious type of error. Cutscores for surgeons or airline pilots can therefore be adjusted upward to reduce the possibility that incompetent individuals will pass the test. Note, however, that this greater certainty is purchased at the price of an increased proportion of truly competent individuals who will *fail* the test.

In some other cases, "false negative" errors may be judged to be the more serious type because of the harm to individuals who are kept out of a profession or the harm to society if qualified candidates fail the test. In minimum competency testing programs, for example, the passing score may be adjusted *downward* to reduce the possibility that some students who are competent will nonetheless fail.

Impact of Method on Results

The most consistent finding from the research literature on standard setting is that different methods lead to different results. Not only do judgmental and empirical methods lead to different results, as might be expected, but different judgmental methods lead to different results. In fact, judgmental methods appear to be sensitive to slight and seemingly trivial differences in the procedures used to implement a given method. Jaeger attempted to quantify the differences in standards set within a single study, using different methods but keeping the test and judges the same.[7] He found that, in such studies, standards set by different methods on the same test typically produced differences in passing rates that were three to six times as great from one to the next. For example, if the most rigorous standard in a study passed 5 percent of the examinees, the other standard passed 15 percent or 30 percent. Such differences are large and practically significant. Also recall that these studies were carried out with a single group of judges or with randomly equivalent groups of judges. Real differences in judges' opinions and value perspectives would create *additional* variation in proposed standards. This might occur, for example, if subject-matter experts and lay groups were invited to deliberate independently to recommend standards.

[7] Ibid.

Standard Setting as Value Judgments

Because standard-setting methods were developed in the psychometric literature and involve numerical calculations, they are often imbued with a false sense of scientific precision. In reality, measurement experts acknowledge that no true standards exist in nature for standard-setting techniques to "estimate." Therefore, a standard-setting method, even when implemented precisely as recommended in the literature, will not necessarily produce a true or valid standard. It will only summarize the judges' opinions. A better way to think about standard-setting methods is that they are strategies to help judges think more systematically about the judgments they must make. Livingston, who authored *Passing Scores,* a standard reference in the field, has commented more recently that the terminology "standard setting methods" might better be replaced by "standard setting studies." According to Livingston, such studies are not intended to produce a single cutscore number but only to serve as one source of information. Ultimately "the choice of a standard is a policy decision." [8]

Given that no one standard-setting method produces "truth," it has been suggested that more than one source of relevant information be considered before arriving at a final standard. For example, one might consider the *range* of cutscores produced by the various individual judges (not just their average recommendation), along with external validity evidence such as predictive relationships or concurrent performance of examinee groups outside the test situation.

NAGB's Rationale for Using the Modified Angoff Method

The psychometric literature was reviewed by NAGB staff; they were aware that the various judgmental and empirical methods often produce different results. They explained their recommendation for using the Angoff procedure as follows:

> First, the advantages and disadvantages of many of the competing procedures are well documented in the literature. There have been any number of research studies completed documenting some of the differences; the Angoff procedure is generally superior. Secondly, it is quite straightforward; both the judging task and its results are intuitively interpretable. Thirdly, it does not require the administration of items to a trial population. This means, of course, that standard setting could begin in the immediate future.[9]

Originally NAGB staff recommended setting a single standard for each grade to reflect the "core of learning in each field that every student ought to master." [10] In addition,

[8] Samuel A. Livingston, personal communication with the author, April 29, 1993.

[9] Roy Truby, "Staff Paper on Setting Goals for the National Assessment" (Washington, D.C.: National Assessment Governing Board, December 8, 1989), 15.

[10] Ibid., 9.

a second standard would be set at 12th grade for those students going into college work that would require higher-level courses. Eventually, when the decision was made to specify three levels of achievement (as well as an implied below basic level), the Angoff procedure was modified to facilitate making judgments about three different cutpoints for each grade. More information about NAGB's judgmental procedures is reported in a subsequent section.

NAGB staff also reviewed the contrasting-groups method for setting standards, noting that it would take longer to implement because it requires collecting field-test data. The staff concluded that "it would be advisable for NAEP to use both of these procedures: judgment methods to establish standards, and empirical methods as a verification procedure." [11]

Previous Evaluations of the NAEP Achievement Levels

Earlier evaluations focused on the validity of the 1990 mathematics achievement levels. Difficulties identified in these studies led to changes in NAGB's procedures but also to calls for further evaluation. As background to the present report, which focuses on the 1992 procedures and achievement levels, we provide here a brief summary of previous evaluations. We also include key findings from the just-released General Accounting Office (GAO) evaluation, which was conducted concurrently with the Panel's evaluation and encompasses both 1990 and 1992.

As noted in chapter 1, the Panel's current evaluation is independent of these previous evaluations. However, to the extent that the Panel's findings corroborate the earlier findings, both are strengthened. Some specific points of agreement between the Panel and earlier evaluations will be discussed in the substantive chapters that follow.

Results of Evaluation by the Technical Review Panel

A Technical Review Panel (TRP) was established by NCES to provide advice regarding the validity of NAEP on an ongoing basis. In 1990 a subgroup of the TRP was charged with evaluating NAGB's initial effort to set achievement levels.[12] Analyses showed large inconsistencies in levels set by different groups of judges; judges' ratings were incoherent given that 8th graders were expected to outperform 12th graders; and the advanced-level cutpoints were set in regions of the scale that could not be measured accurately because there were insufficient numbers of challenging items on the test. The TRP also commented that there was no evidence to support the predictive claims conveyed in the achievement-level descriptions (e.g., that attainment at the proficient level represented adequate preparation for the next level of schooling). The TRP concluded that the achievement levels were "seriously flawed" and recommended that

[11] Ibid., 18.

[12] R.L. Linn, D.M. Koretz, E.L. Baker, and L. Burstein, *The Validity and Credibility of the Achievement Levels for the 1990 National Assessment of Educational Progress in Mathematics* (Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, January 1991).

they not be used in any public reporting of NAEP results. In response to concerns raised by the TRP and by its own contractor and evaluators, NAGB held additional rating sessions and revised the achievement levels before releasing its 1990 report.

Results of the Evaluation Commissioned by NAGB

NAGB itself commissioned an evaluation of the 1990 achievement levels by Stufflebeam, Jaeger, and Scriven.[13] The Stufflebeam et al. report considered both the first and second phases of the level-setting process and was severely critical of the final achievement levels. The evaluators recommended that the proposed levels not be used as a baseline to measure future changes in mathematics achievement, and they recommended against the use of the same procedures to set levels for future NAEP assessments. Stufflebeam and his colleagues acknowledged the conscientious efforts made by NAGB to improve the process within the constraints of severe time pressures, but the evaluation team still identified five major problems: (1) judges were confused by the definitions and their task (e.g., instructions to judges required them to conceive of a double-hypothetical—how a hypothetical borderline student might, hypothetically, perform on each of a set of test items); (2) the item pool was not adequate, especially to represent advanced performance; (3) procedures were chaotic at first, and the project had to be recycled; (4) technical flaws such as the mismatch between end-of-year standards and mid-year testing and instructions about guessing were never resolved; and (5) although the final phase-two achievement levels (derived by averaging across four regional panels of judges) showed a more coherent progression across grades than the rejected phase-one achievement levels, the differences between the four separate panels of judges were large, indicating that the standards were highly dependent on the particular panel of judges.

In spite of the critical recommendations contained in the Stufflebeam et al. report, NAGB decided to report the 1990 achievement levels as a trial effort. While the provisional nature of the achievement levels was not heavily emphasized in the public reporting, it was noted. For example, the NAGB report publishing the 1990 results by level is subtitled *Initial Performance Standards for the 1990 NAEP Mathematics Assessment,* and the executive summary concludes with this paragraph:

> The development and application of performance level standards represents an initial effort. These processes have been, and will continue to be, carefully evaluated by the Board and others. The Board remains committed to the use of performance level standards and will be continuing these activities in connection with future administrations of NAEP, including the assessments of mathematics, writing, and reading scheduled for 1992.[14]

Also, in the press conference held to present the achievement levels to the public on September 30, 1991, attendees were told that the achievement-level-setting process

[13] D. Stufflebeam, R.M. Jaeger, and M. Scriven, *Summative Evaluation of the National Assessment Governing Board's Inaugural 1990-91 Effort to Set Achievement Levels on the National Assessment of Educational Progress* (Washington, D.C.: National Assessment Governing Board, August 1991).

[14] M.L. Bourque and H. Garrison, *The LEVELS of Mathematics Achievement: Initial Performance Standards for the 1990 NAEP Mathematics Assessment* (Washington, D.C.: National Assessment Governing Board, 1991), ix.

would be repeated in 1992, but that the 1990 levels were being released in order to stimulate public discussion and determine how useful the levels proved to be as a reporting device.

NAE Panel's First Report

Additional criticisms continued to be leveled against the release of the 1990 achievement levels and the adequacy of NAGB's achievement-level-setting process. In the NAE Panel's report on the 1990 Trial State Assessment, it was shown that reporting by achievement levels or anchor points would give virtually the same state-by-state results in eighth-grade mathematics. This occurred because the basic cutpoint (255) was so close to the 250 anchor point and the proficient cutpoint (295) was very close to the 300 anchor point. Therefore, based on its appraisal of other evaluation studies, the Panel made the following recommendations: (1) The 1990 achievement levels should not be used as a baseline for reporting future NAEP results; (2) anchor points should continue to be reported even if achievement-level results are also released; (3) the 1992 mathematics assessment should be reported by subscore, not just by total NAEP score (achievement levels were set only for the total score); (4) NAGB and the National Education Goals Panel should agree on a single reporting scheme to avoid confusion in the future;[15] (5) NAGB should conduct studies to validate the descriptors used to illustrate the achievement levels—especially those implying outcomes of future performance; and (6) the procedures for setting the 1992 achievement levels should be evaluated by a distinguished team of external evaluators as well as by an internal team. The present evaluation was commissioned by NCES in response to this last recommendation.

Results of Evaluation by the General Accounting Office

Concurrent with the Panel's work, the U.S. House of Representatives Committee on Education and Labor requested that the Program Evaluation and Methodology Division of the GAO review the methodology used by NAGB to set achievement levels. The Committee was concerned with the technical quality of the procedures because they might affect interpretations of NAEP data, but also because NAGB's procedures might be adopted by other organizations such as the Goals Panel.[16]

Interim report. In their interim report, issued in March 1992, the GAO addressed some policy and governance issues such as the proper role of NAGB, a lay body, in

[15] The Goals Panel released the same 1990 assessment results in mathematics in a form different from NAGB's. The advanced and proficient levels were combined into a single "competent" category; basic and below basic became "not competent." See National Education Goals Panel, *The National Education Goals Report: Building a Nation of Learners* (Washington, D.C.: Author, 1991), 220.

[16] Honorable William D. Ford, Chairman, Committee on Education and Labor, House of Representatives, and Honorable Dale E. Kildee, Chairman, Subcommittee on Elementary, Secondary, and Vocational Education and Labor, House of Representatives, personal communication to Eleanor Chelimsky at U.S. General Accounting Office, Washington, D.C., October 7, 1991.

making technical decisions.[17] On the technical side, GAO findings were very similar to other evaluations of the 1990 levels. They noted the unreliability of judges' decisions and the lack of empirical validity data. As a rough approximation of the type of validity evidence needed, the GAO examined the proportions of examinees reaching the levels among groups of disadvantaged and high-ability students. The levels appeared to be set unrealistically high. In fourth-grade high-ability classes, for example, fewer than 5 percent reached the advanced level while more than 10 percent failed to reach the basic level. According to the GAO, NAGB did a good job of responding to advice about how to improve the item-judgment procedure, but they did not respond to fundamental questions raised about the approach itself. Although aware of the criticisms, "NAGB did not disclose the limitations of the levels data when it published the 1990 mathematics levels results."[18] The GAO concluded by recommending that achievement levels not be used to "organize the reporting and analysis of 1992 NAEP results" because they could not be thoroughly examined in time to meet the publication schedule.[19]

Final report. The final GAO report, issued in July 1993, continued the evaluation to include the 1992 achievement levels.[20] This report concluded that the 1992 procedures, implemented by American College Testing (ACT) under contract to NAGB, addressed some of the problems that affected the 1990 standard setting, but did nothing to resolve the fundamental problem—that the mastery dimension of NAGB's achievement levels cannot be "adequately ... represented using the NAEP test and scale, which were designed to depict overall performance."[21] That is, in the case of mathematics, for example, the mastery of specific mathematical content and processes, as laid out in the mathematics achievement-level descriptions, does not align with any specific cutscore on the overall NAEP mathematics scale. Moreover, some of the mathematical behaviors implied by the achievement levels are not measured in the current NAEP item pool. Consequently, anyone trying to interpret NAEP scores by reference to the achievement-level descriptions provided would misinterpret the capabilities of U.S. students in a number of subtle and not-so-subtle ways. The GAO concludes with a set of recommendations:

> (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.[22]

The GAO also noted that NAGB's standard-setting procedure was applied to reading and writing for the first time in 1992. This required adapting the Angoff procedure to allow judges to estimate cutscores associated with extended-response questions on which examinees could receive either full or partial credit. "ACT pilot-tested procedures for judging such items and concluded that they were feasible, but it is too

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

[18] Ibid., 4.

[19] Ibid., 5.

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

[21] Ibid., 34.

[22] Ibid., 5.

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."[23, 24]

The 1992 Procedures for Setting Achievement Levels

The foregoing evaluation studies primarily addressed the 1990 achievement-level procedures and results. In response, NAGB made substantial efforts to improve the consistency and coherence of the Angoff procedure for implementation with the 1992 assessments, as well as directing the NAEP contractor to make some modifications to the NAEP item pools. However, because of its commitment to make achievement-level standards available for the 1992 assessment, NAGB did not attend to advice that would have implied slowing down the schedule (to gather empirical evidence) or abandoning the Angoff effort (in favor of other approaches).

The Panel's present studies address the more recent standard-setting efforts. A brief outline of the current procedures is provided here; additional methodological details are given in subsequent chapters when they are relevant to the Panel's analyses.[25]

Selecting and Training Panelists

NAGB developed a request for proposals and contracted with ACT to convene the level-setting panels for the 1992 assessment. ACT is a well-recognized testing firm and has had considerable experience with setting passing scores for professional licensure exams. Details of the procedures (e.g., selection of judges, training of judges) were specified in advance and reviewed by advisory committees. A pilot study was conducted in reading to evaluate the training materials and logistical procedures.

The approach to selecting judges for 1992 ensured that the panels were broadly representative. Nominations of individuals knowledgeable in each subject area were sought from various professional and political organizations. From the nominees, panelists were selected to represent teachers, nonteacher educators, and noneducators, and to be balanced with respect to gender, ethnicity, and geographic region. For each subject area, between 20 and 24 panelists per grade level participated.

Panel members were given background materials to read prior to the meeting. The three-and-one-half-day level-setting sessions in each topic area (mathematics

[23] Ibid., 36.

[24] After reviewing the results of their level-setting process, NAGB decided not to report the 1992 writing results in terms of achievement levels. The reading achievement levels will be used in reporting.

[25] This summary is based on: American College Testing Program, *Descriptions of Mathematics Achievement Levels-Setting Process and Proposed Achievement Level Definitions* (Iowa City, IA: Author, April 13, 1992), and M.L. Bourque, "The NAEP Achievement Level Setting Process for the 1992 Mathematics Assessment," In *Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics,* E.G. Johnson, J. Mazzeo, and D.L. Kline (Washington, D.C.: National Center for Education Statistics, 1993).

and reading) began with an orientation to NAEP and NAGB, the NAEP frameworks, and the generic NAGB definitions of the achievement levels. Panelists next took and scored a grade-appropriate NAEP booklet. The purpose of this exercise was to familiarize judges with NAEP content and scoring rubrics before the development of grade- and subject-specific descriptions for the levels. (The items were subsequently redistributed among panelists so that any individual panelist would rate different items than he or she had used in training.)

Developing Achievement-Level Descriptions

Panelists worked in small groups to generate lists describing what students should be able to do at each grade level to be considered basic, proficient, or advanced in mathematics or reading (based on NAGB's generic descriptions of these levels). NAEP frameworks and some NAEP test items were available to the panelists during this process, but the resultant achievement levels were not exclusively reflective of either of these sources.

After discussion to identify those elements that best exemplified each achievement level, panelists agreed on a final list of descriptors. The descriptors were synthesized into descriptions for the achievement levels which were used as a basis for subsequent item-by-item judgments about what students should be able to do at each achievement level. Because the level descriptions in both reading and mathematics were revised subsequently, the descriptions developed and used at the initial level-setting meetings are referred to in this report as the St. Louis descriptions (after the city where the level-setting meetings were held).

Defining Cutscores and Selecting Exemplar Items

Next, panelists made judgments about test content to identify specific task and item performances that exemplified each level. In reading and mathematics, panelists were given 2 hours of training in the modified Angoff method and then participated in three rounds of item ratings as follows.[26] (The procedure is referred to as a "modified" Angoff approach because it was done iteratively with feedback after each round and because the judges set three levels instead of one.) In Round 1, panelists first answered the items and scored them and then, item by item, made three ratings corresponding to the three achievement-level cutpoints. Using their descriptions, panelists were to estimate the probability that a "borderline" examinee would get the item correct for each of the basic, proficient, and advanced levels. In Round 2, judges rerated the same items after being provided with data on the item difficulties (p-values) and the consistency of their ratings with those of other judges. Before Round 3, judges were given additional information about their own internal consistency (i.e., whether they were estimating similar probabilities for items of similar difficulty).

[26] In writing, as well as for the extended-response items in mathematics and reading, the process did not follow the Angoff methodology; instead judges were asked to identify two papers that represented borderline performance for each of the levels.

On the last day, after the judges' average ratings had been translated into cutpoints, the panelists were given sample NAEP items that 50 percent of the borderline students at each level would get correct, and were asked to select from them exemplar items that were consistent with the level descriptions. However, it was later concluded that exemplar items should represent performance typical of the *entire level* rather than borderline performance, so these exemplars were not used. Instead, different sets of exemplar items were selected at the subsequent validation meetings.

Convening Followup Validation Panels

Validation panels were convened in both reading and mathematics to review the level descriptions and select exemplar items. Participants at these meetings included some of the teachers who had participated at the initial level-setting meeting, plus content experts. For example, in mathematics, the new panelists included representatives from the National Council of Teachers of Mathematics, the Mathematical Sciences Education Board, and state mathematics curriculum supervisors.[27] Panelists were asked to review and revise the descriptions to ensure consistency with the NAGB generic definitions and the NAEP frameworks, appropriateness for the target grade, within- and across-grade consistency, alignment with professional content standards, and "utility for increasing the public's understanding of the NAEP [mathematics] results."[28] Members of the validation panel were also asked to review possible exemplar items and to select items for each level "that would best communicate to the public the levels of [mathematics] ability and the types of skills needed to perform [in mathematics] at that level."[29]

Adopting Final Cutscores

To set the final cutpoints in mathematics, NAGB took the average Round 3 rating for each level, subtracted one standard error of measurement (based on split-sample variation in judges' ratings), and adopted these values as the final cutpoints. In reading, the cutpoints recommended from the Angoff process were adopted directly without further adjustment. Since the 1992 mathematics standards were not the same as those set in 1990, the 1992 standards were also applied retrospectively to determine the proportions of 1990 students testing above each level. This was necessary in order to permit comparisons of student achievement across the two years. (See table 2.2.)

[27] Bourque, op. cit.

[28] Ibid., 376.

[29] Ibid., brackets added.

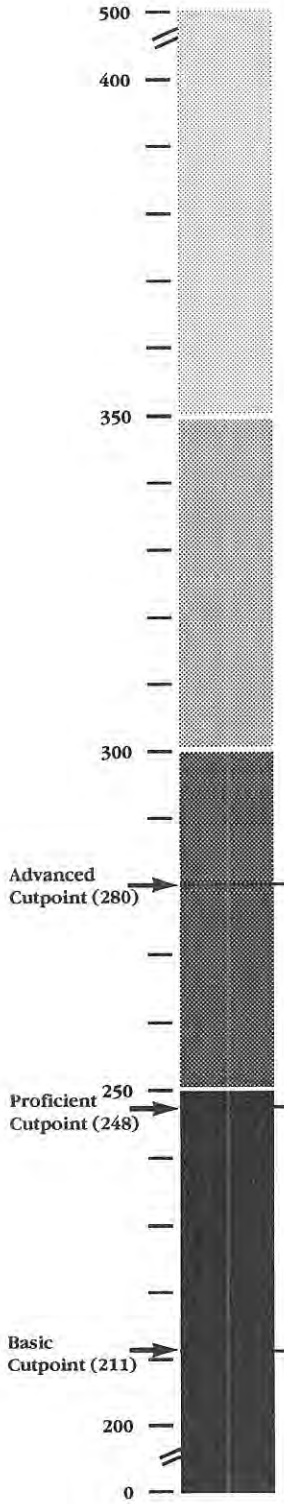
Reporting of the 1992 Achievement Levels in Mathematics and Reading

The 1992 achievement levels for each grade and subject are presented in figures 2.2 to 2.7. Each figure shows the numeric achievement-level cutpoints, as well as the verbal descriptions intended to characterize performance at the achievement levels. As noted above, these descriptions were developed independently of the test items and do not necessarily reflect the specific capabilities of students whose overall performance would place them at a given achievement level. Moreover, the exemplary items published to illustrate the achievement levels did not necessarily have performance characteristics that would anchor them in the scale range associated with the relevant achievement level. This discrepancy was strongly criticized by a number of parties, including the GAO, the TRP, and persons involved in the NCES internal report adjudication process, when it became apparent during the review of the draft 1992 mathematics reports.[30] The response from NAGB was that the achievement-level descriptions were meant to portray more general aspirations for student achievement and should not be narrowly tied to the content of a particular (and imperfect) test.

After some delay, NCES acceded to the concerns of user groups who were anxious to obtain the 1992 state-by-state mathematics results, and published the full national and state mathematics reports with some explanatory caveats. A more satisfactory compromise was reached for the 1992 reading reports, expected to be released in September 1993. For these reports, ETS applied a behavioral anchoring technique similar to that which has traditionally been used for the anchor points, to generate post-hoc, empirical descriptions of the skills and behaviors exhibited by students whose overall performance places them at a particular achievement level. These descriptions will be reported along with the more prescriptive achievement-level descriptions prepared by NAGB. The report will explain that the descriptions developed through behavioral anchoring describe what students scoring at a given level *can* do (as measured by the current test), while the achievement-level descriptions present an education standard for what students *should* do in order to earn the appellations of basic, proficient, and advanced.

[30] U.S. General Accounting Office, op. cit. (1993); Technical Review Panel, personal communication to Gary Phillips, Washington, D.C., March 1993.

Figure 2.2. ***Grade 4 math achievement-level cutpoints and descriptions***



ADVANCED: Fourth-grade students performing at the advanced level should apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problem solving in the five NAEP content areas.

Fourth graders performing at the advanced level should be able to solve complex and nonroutine real-world problems in all NAEP content areas. They should display mastery in the use of four-function calculators, rulers, and geometric shapes. These students are expected to draw logical conclusions and justify answers and solution processes by explaining why, as well as how, they were achieved. They should go beyond the obvious in their interpretations and be able to communicate their thoughts clearly and concisely.

PROFICIENT: Fourth-grade students performing at the proficient level should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content areas.

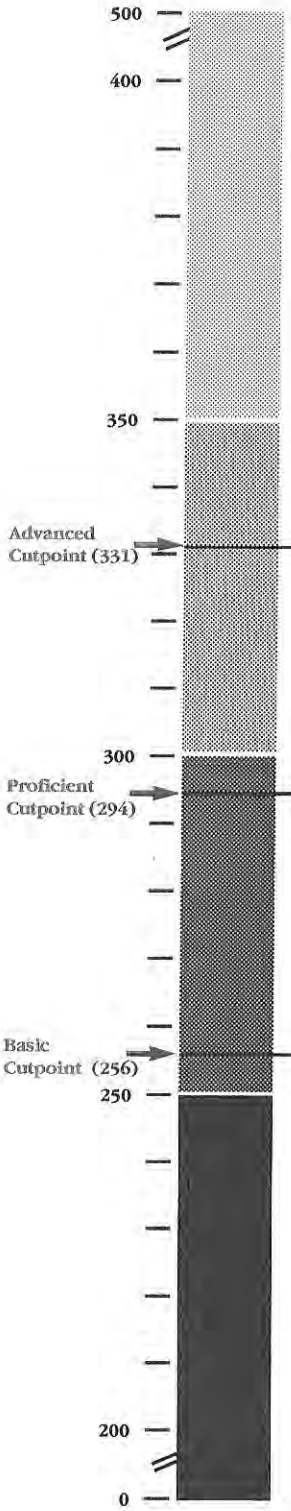
Fourth graders performing at the proficient level should be able to use whole numbers to estimate, compute, and determine whether results are reasonable. They should have a conceptual understanding of fractions and decimals; be able to solve real-world problems in all NAEP content areas; and use four-function calculators, rulers, and geometric shapes appropriately. Students performing at the proficient level should employ problem-solving strategies such as identifying and using appropriate information. Their written solutions should be organized and presented both with supporting information and explanations of how they were achieved.

BASIC: Fourth-grade students performing at the basic level should show some evidence of understanding the mathematical concepts and procedures in the five NAEP content areas.

Fourth graders performing at the basic level should be able to estimate and use basic facts to perform simple computations with whole numbers; show some understanding of fractions and decimals; and solve some simple real world problems in all NAEP content areas. Students at this level should be able to use–though not always accurately–four-function calculators, rulers, and geometric shapes. Their written responses are often minimal and presented without supporting information.

BELOW BASIC: This category has no formal description.

Figure 2.3. ***Grade 8 math achievement-level cutpoints and descriptions***



ADVANCED: Eighth-grade students performing at the advanced level should be able to reach beyond the recognition, identification, and application of mathematical rules in order to generalize and synthesize concepts and principles in the five NAEP content areas.

Eighth graders performing at the advanced level should be able to probe examples and counterexamples in order to shape generalizations from which they can develop models. Eighth graders performing at the advanced level should use number sense and geometric awareness to consider the reasonableness of an answer. They are expected to use abstract thinking to create unique problem-solving techniques and explain the reasoning processes underlying their conclusions.

PROFICIENT: Eighth-grade students performing at the proficient level should apply mathematical concepts and procedures consistently to complex problems in the five NAEP content areas.

Eighth graders performing at the proficient level should be able to conjecture, defend their ideas, and give supporting examples. They should understand the connections between fractions, percents, decimals, and other mathematical topics such as algebra and functions. Students at this level are expected to have a thorough understanding of basic level arithmetic operations—an understanding sufficient for problem solving in practical situations.

Quantity and spatial relationships in problem solving and reasoning should be familiar to them, and they should be able to convey underlying reasoning skills beyond the level of arithmetic. They should be able to compare and contrast mathematical ideas and generate their own examples. These students should make inferences from data and graphs; apply properties of informal geometry; and accurately use the tools of technology. Students at this level should understand the process of gathering and organizing data and be able to calculate, evaluate, and communicate results within the domain of statistics and probability.

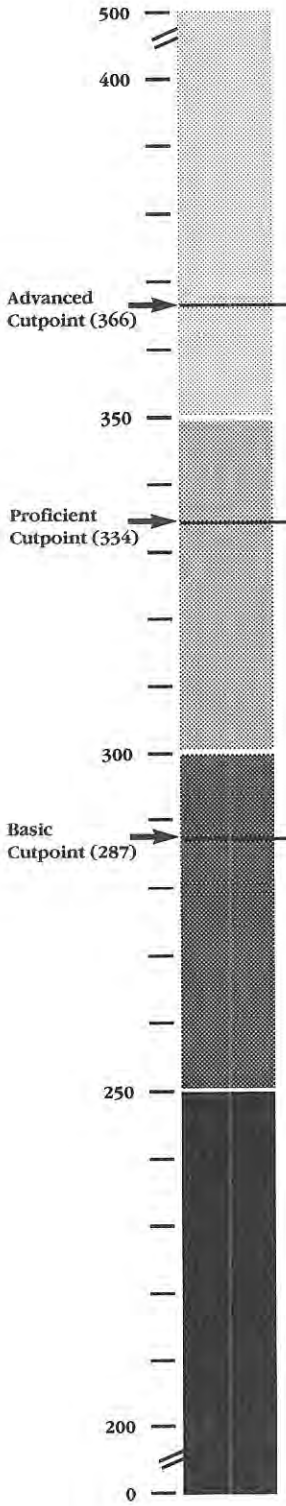
BASIC: Eighth-grade students performing at the basic level should exhibit evidence of conceptual and procedural understanding in the five NAEP content areas. This level of performance signifies an understanding of arithmetic operations—including estimation—on whole numbers, decimals, fractions, and percents.

Eighth graders performing at the basic level should complete problems correctly with the help of structural prompts such as diagrams, charts, and graphs. They should be able to solve problems in all NAEP content areas through the appropriate selection and use of strategies and technological tools—including calculators, computers, and geometric shapes. Students at this level also should be able to use fundamental algebraic and informal geometric concepts in problem solving.

As they approach the proficient level, students at the basic level should be able to determine which of available data are necessary and sufficient for correct solutions and use them in problem solving. However, these eight graders show limited skill in communicating mathematically.

BELOW BASIC: This category has no formal description.

Figure 2.4. *Grade 12 math achievement-level cutpoints and descriptions*



ADVANCED: Twelfth-grade students performing at the advanced level should consistently demonstrate the integration of procedural and conceptual knowledge and the synthesis of ideas in the five NAEP content areas.

Twelfth-grade students performing at the advanced level should understand the function concept; and be able to compare and apply the numeric, algebraic, and graphical properties of functions. They should apply their knowledge of algebra, geometry, and statistics to solve problems in more advanced areas of continuous and discrete mathematics.

They should be able to formulate generalizations and create models through probing examples and counterexamples. They should be able to communicate their mathematical reasoning through the clear, concise, and correct use of mathematical symbolism and logical thinking.

PROFICIENT: Twelfth-grade students performing at the proficient level should consistently integrate mathematical concepts and procedures to the solutions of more complex problems in the five NAEP content areas.

Twelfth graders performing at the proficient level should demonstrate an understanding of algebraic, statistical, and geometric and spatial reasoning. They should be able to perform algebraic operations involving polynomials; justify geometric relationships; and judge and defend the reasonableness of answers as applied to real-world situations. These students should be able to analyze and interpret data in tabular and graphical form; understand and use elements of the function concept in symbolic, graphical, and tabular form; and make conjectures, defend ideas, and give supporting examples.

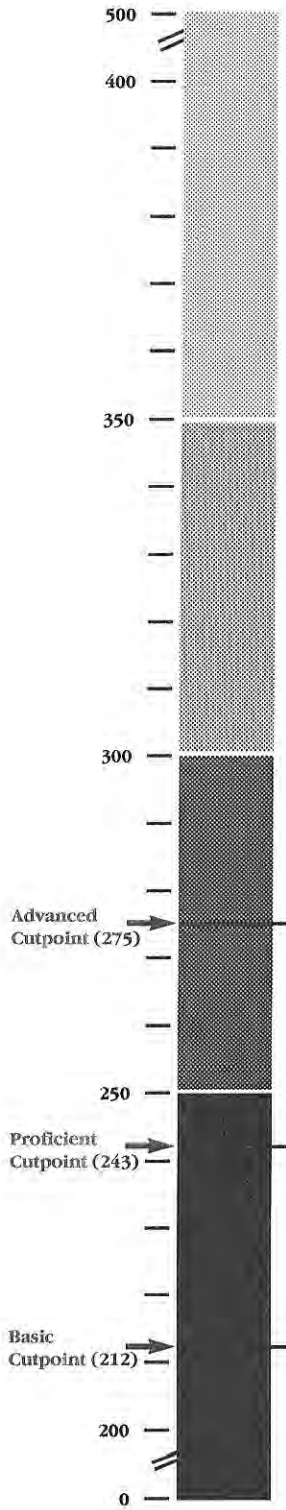
BASIC: Twelfth-grade students performing at the basic level should demonstrate procedural and conceptual knowledge in solving problems in the five NAEP content areas.

Twelfth-grade students performing at the basic level should be able to use estimation to verify solutions and determine the reasonableness of results as applied to real-world problems. They are expected to use algebraic and geometric reasoning strategies to solve problems. Twelfth graders performing at the basic level should recognize relationships presented in verbal, algebraic, tabular, and graphical forms; and demonstrate knowledge of geometric relationships and corresponding measurement skills.

They should be able to apply statistical reasoning in the organization and display of data and in reading tables and graphs. They also should be able to generalize from patterns and examples in the areas of algebra, geometry, and statistics. At this level, they should use correct mathematical language and symbols to communicate mathematical relationships and reasoning processes; and use calculators appropriately to solve problems.

BELOW BASIC: This category has no formal description.

Figure 2.5. ***Grade 4 reading achievement-level cutpoints and descriptions***



ADVANCED: Fourth-grade students performing at the advanced level should be able to generalize about topics in the reading selection and demonstrate an awareness of how authors compose and use literary devices. When reading text appropriate to fourth grade, they should be able to judge texts critically and, in general, give thorough answers that indicate careful thought.

For example, when reading literary text, advanced-level students should be able to make generalizations about the point of the story and extend its meaning by integrating personal experiences and other readings with the ideas suggested by the text. They should be able to identify literary devices such as figurative language.

When reading informational text, advanced-level fourth graders should be able to explain the author's intent by using supporting material from the text. They should be able to make critical judgments of the form and content of the text and explain their judgments clearly.

PROFICIENT: Fourth-grade students performing at the proficient level should be able to demonstrate an overall understanding of the text, providing inferential as well as literal information. When reading text appropriate to fourth grade they should be able to extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences. The connection between the text and what the student infers should be clear.

For example, when reading literary text, proficient-level fourth graders should be able to summarize the story, draw conclusions about the characters or plot, and recognize relationships such as cause and effect.

When reading informational text, proficient-level students should be able to summarize the information and identify the author's intent or purpose. They should be able to draw reasonable conclusions from the text, recognize relationships such as cause and effect or similarities and differences, and identify the meaning of the selection's key concepts.

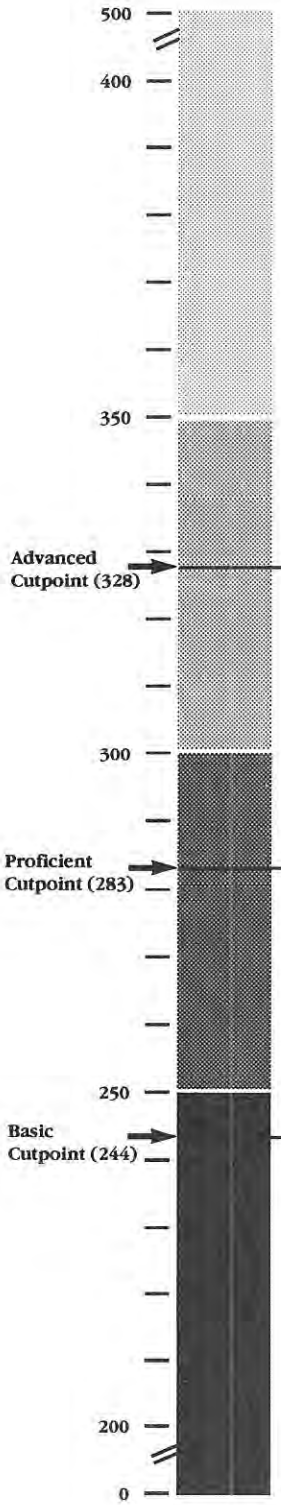
BASIC: Fourth-grade students performing at the basic level should demonstrate an understanding of the overall meaning of what they read. When reading texts appropriate for fourth graders, they should be able to make relatively obvious connections between the text and their own experiences.

For example, when reading literary text, they should be able to tell what the story is generally about—providing details to support their understanding—and be able to connect aspects of the stories to their own experiences.

When reading informational text, basic-level fourth graders should be able to tell what the selection is generally about or identify the purpose for reading it; provide details to support their understanding; and connect ideas from the text to their background knowledge and experiences.

BELOW BASIC: This category has no formal description.

Figure 2.6. *Grade 8 reading achievement-level cutpoints and descriptions*



ADVANCED: Eighth-grade students performing at the advanced level should be able to describe the more abstract themes and ideas of the overall text. When reading text appropriate to eighth grade, they should be able to analyze both meaning and form and support their analyses explicitly with examples from the text; they should be able to extend text information by relating it to their experiences and to world events. At this level, student responses should be thorough, thoughtful, and extensive.

For example, when reading literary text, advanced-level eighth graders should be able to make complex, abstract summaries and theme statements. They should be able to describe the interactions of various literary elements (i.e., setting, plot, characters, and theme); to explain how the use of literary devices affects both the meaning of the text and their response to the author's style. They should be able critically to analyze and evaluate the composition of the text.

When reading informative text, they should be able to analyze the author's purpose and point of view. They should be able to use cultural and historical background information to develop perspectives on the text and be able to apply text information to broad issues and world situations.

When reading practical text, advanced-level students should be able to synthesize information that will guide their performance, apply text information to new situations, and critique the usefulness of the form and content.

PROFICIENT: Eighth-grade students performing at the proficient level should be able to show an overall understanding of the text, including inferential as well as literal information. When reading text appropriate to eighth grade, they should extend the ideas in the text by making clear inferences from it, by drawing conclusions, and by making connections to their own experiences—including other reading experiences. Proficient eighth graders should be able to identify some of the devices authors use in composing text.

For example, when reading literary text, students at the proficient level should be able to give details and examples to support themes that they identify. They should be able to use implied as well as explicit information in articulating themes; to interpret the actions, behaviors, and motives of characters; and to identify the use of literary devices such as personification and foreshadowing.

When reading informative text, they should be able to summarize the text using explicit and implied information and support conclusions with inferences based on the text.

BASIC: Eighth-grade students performing at the basic level should demonstrate a literal understanding of what they read and be able to make some interpretations. When reading text appropriate to eighth grade, they should be able to identify specific aspects of the text that reflect the overall meaning, recognize and relate interpretations and connections among ideas in the text to personal experience, and draw conclusions based on the text.

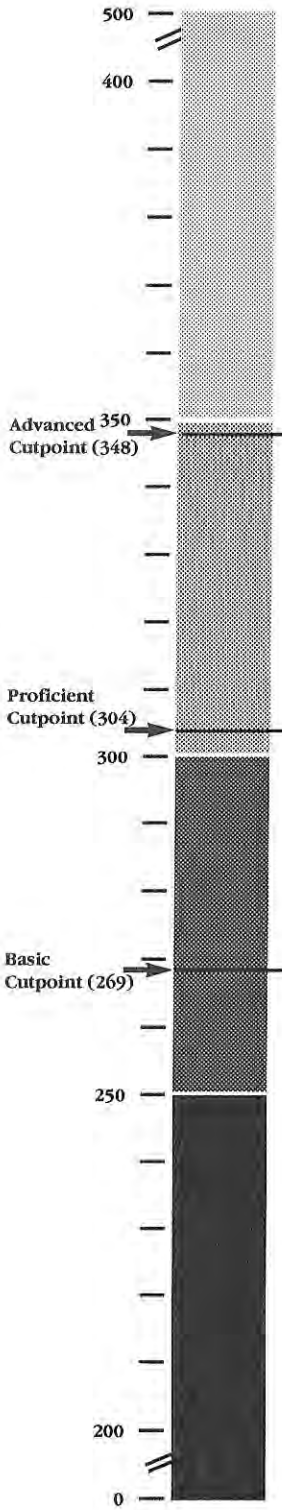
For example, when reading literary text, basic-level eighth graders should be able to identify themes and make inferences and logical predicitions about aspects such as plot and characters.

When reading informative text, they should be able to identify the main idea and the author's purpose. They should make inferences and draw conclusions supported by information in the text. They should recognize the relationships among the facts, ideas, events, and concepts of the text (e.g., cause and effect and chronological order).

When reading practical text, they should be able to identify the main purpose and make predictions about the relatively obvious outcomes of procedures in the text.

BELOW BASIC: This category has no formal description.

Figure 2.7. ***Grade 12 reading achievement-level cutpoints and descriptions***



ADVANCED: Twelfth-grade students performing at the advanced level should be able to describe more abstract themes and ideas in the overall text. When reading text appropriate to twelfth grade, they should be able to analyze both the meaning and the form of the text and explicitly support their analyses with specific examples from the text. They should be able to extend the information from the text by relating it to their experiences and to the world. Their responses should be thorough, thoughtful, and extensive.

For example, when reading literary text, advanced-level twelfth graders should be able to produce complex, abstract summaries and theme statements. They should be able to use cultural, historical, and personal information to develop and explain text perspectives and conclusions. They should be able to evaluate the text, applying knowledge gained from other texts.

When reading informational text, they should be able to analyze, synthesize, and evaluate points of view. They should be able to identify the relationship between the author's stance and elements of the text. They should be able to apply text information to new situations and to the process of forming new responses to problems or issues.

When reading practical text, advanced-level twelfth graders should be able to make a critical evaluation of the usefulness of the text and apply directions from the text to new situations.

PROFICIENT: Twelfth-grade students performing at the proficient level should be able to show an overall understanding of the text which includes inferential as well as literal information. When reading text appropriate to twelfth grade, they should be able to extend the ideas of the text by making inferences, drawing conclusions, and making connections to their own personal experiences and other readings. Connections between inferences and the text should be clear, even when implicit. These students should be able to analyze the author's use of literary devices.

When reading literary text, proficient-level twelfth graders should be able to integrate their personal experiences with ideas in the text to draw and support conclusions. They should be able to explain the author's use of literary devices such as irony or symbolism.

When reading informative text, they should be able to apply text information appropriately to specific situations and integrate their background information with ideas in the text to draw and support conclusions.

When reading practical text, they should be able to apply information or directions appropriately. They should be able to use personal experiences to evaluate the usefulness of text information. When reading practical text, proficient-level students should be able to describe its purpose and support their views with examples and details. They should be able to judge the importance of certain steps and procedures.

BASIC: Twelfth-grade students performing at the basic level should be able to demonstrate an overall understanding and make some interpretations of the text. When reading text appropriate to twelfth grade, they should be able to identify and relate aspects of the text to its overall meaning, recognize interpretations, make connections among and relate ideas in the text to their personal experiences, and draw conclusions. They should be able to identify elements of an author's style.

For example, when reading literary text, twelfth-grade students should be able to explain the theme, support their conclusions with information from the text, and make connections between aspects of the text and their own experiences.

When reading informational text, basic-level twelfth graders should be able to explain the main idea or purpose of a selection and use text information to support a conclusion or make a point. They should be able to make logical connections between the ideas in the text and their own background knowledge.

When reading practical text, they should be able to explain its purpose and the significance of specific details or steps.

BELOW BASIC: This category has no formal description.

Summary

This chapter has provided background relevant to the Panel's evaluation of the 1992 achievement levels in reading and mathematics. NAGB's standard-setting efforts are part of a long tradition of refining the reporting of NAEP results in order to improve their interpretability. In particular, there has been a growing interest in adding an evaluative component to NAEP reporting which would facilitate tracking progress toward education goals.

A large number of standard-setting methods are documented in the tests and measurement literature. Some require judges to set cutpoints after closely studying the items on a test, while others require that judges evaluate examinee performance outside of the test context, then set cutscores based on the actual test performance of examinees they have judged to be competent. The test-based methods are commonly called "judgmental methods," and the examinee-based methods are called "empirical methods." In addition, the literature includes numerous suggestions for methods that allow judges to adjust initial cutscores based on normative information (for example, the proportion of examinees that pass the cutscore compared to the proportion of persons in the population estimated to possess the requisite competence) or some other basis.

One consistent finding in the literature is that different standard-setting methods produce different results. Judgmental methods, in particular, appear to be sensitive to slight and seemingly trivial differences in procedures. Consequently, it may be most appropriate to consider a variety of sources of information before arriving at a final standard.

For the purpose of setting achievement levels on NAEP, NAGB selected the Angoff method, the most widely used and straightforward of the judgmental methods. NAGB rejected the use of empirical methods because these would require the administration of items to a trial population and thus delay the immediate implementation of the standards. NAGB did, however, suggest that the contrasting-groups method, one of the best known empirical methods, might be used as a verification procedure, once the achievement levels had been set.

The most notable of the previous evaluations of the NAGB standard-setting project include studies by the TRP, NAGB's own evaluators (Stufflebeam et al.), and the U.S. General Accounting Office (GAO). The first two focused on the 1990 achievement levels while the GAO report spans both the 1990 and 1992 level-setting efforts.

Major criticisms brought by these previous studies include the following:

- The judgment tasks required by the modified Angoff process were difficult and confusing;
- The NAEP item pool was not adequate to reliably estimate performance at the advanced levels;

- The standards set seemed highly dependent upon the particular sample of judges;
- Appropriate evidence for the predictive or concurrent validity of the cutscores was lacking; and
- Neither the descriptions of student competencies nor the exemplar items were appropriate for describing actual student performance at the designated achievement-level cutscores.

All of the evaluation studies concurred that the achievement levels, as constructed, were not appropriate for reporting the NAEP results.

NAGB was responsive to many of the concerns of its evaluators, but remained committed to delivering final achievement levels for use in reporting 1992 results. Consequently, advice that suggested the need for significant additional data collection, or a fundamental rethinking of the achievement-level-setting process, was not followed.

Additional concerns were raised during the prerelease review of the 1992 mathematics reports. These concerns centered on the misalignment between the achievement-level descriptions and exemplar items, on the one hand, and the actual test performance of students, on the other. Responding to pressure to make the 1992 results available in a timely manner, NCES released the mathematics reports, with explanatory caveats, in April 1993. For the 1992 reading reports, scheduled to be published in September 1993, the prescriptive achievement-level descriptions have been supplemented with more empirical descriptions of the skills and behaviors exhibited by students whose overall performance places them at a particular achievement level.

3 Evaluation of the Process for Setting Achievement Levels

The research literature on standard setting, as summarized in chapter 2, shows that the particular procedures followed in setting standards can substantially (and even capriciously) influence the final cutscore. Therefore, examining the process itself is critical to evaluating the validity of the achievement levels.

The process followed to develop achievement levels for NAEP involved two distinct tasks:

- Creating subject- and grade-specific definitions of each level
- Selecting cutscores to distinguish the levels

The steps followed to complete these tasks were reviewed in chapter 2 and are also summarized in figure 3.1.

Evaluation of the Process for Developing Descriptions

NAGB's achievement-level definitions presented in chapter 1 are "generic"—general statements about performance expectations. In order to be used to evaluate performance on a particular assessment, the generic definitions must be translated into specific requirements for each subject area and grade level. Subject-specific definitions, called "descriptions," tell what students must be able to do in reading or in mathematics for their performance to be considered advanced, proficient, or basic at a given grade. For example, to be advanced in reading at grade 12, students "should be able to describe more abstract themes and ideas in the overall text, ...should be able to analyze both the meaning and the form of the text and explicitly support their analyses with specific examples from the text," and so forth. The first task for panelists at each of the St. Louis meetings was to translate the generic definitions of the achievement levels into specific expectations for their subject area in grades 4, 8, and 12.

For the process to work properly, it was assumed that panelists would develop descriptions in terms of the NAEP frameworks and the content measured by the assessment; otherwise, performance on NAEP could not validly be used to infer progress toward attaining the standards. It was also assumed that the descriptions would form the basis for the subsequent task of deciding on cutscores. After the initial level-setting meetings, the descriptions were revised at subsequent "validation" meetings. Then the descriptions were edited and put in final form by NAGB staff and consultants. (Because the staff version of the descriptions in reading [the third] was judged by content experts to be substantially different from the previous two versions, this report refers to three versions of the descriptions in reading and to only two versions in mathematics.)

Figure 3.1. ***Chronology of steps in the level-setting process***

Development of Subject-Specific Descriptions	Development of Cutscores Using the Modified Angoff Method
1. St. Louis panelists translated NAGB generic achievement-level definitions into subject-specific descriptions for each grade.	
	2. Round 1, based on descriptions, panelists estimated proportion-correct values for right-wrong items at the lower boundary of *basic, proficient,* and *advanced* and selected boundary exemplars for extended responses.
	After each round, proportion-correct values and boundary exemplar scores for all panelists were translated statistically into cutpoints for the lower boundary of each of the levels.
	3. Round 2, panelists received feedback about interjudge consistency, grade-to-grade patterns, and item difficulty, and re-appraised proportion-correct values and boundary exemplars.
	4. Round 3, panelists received feedback about intrajudge and interjudge consistency and grade-to-grade patterns and re-estimated proportion-correct values and boundary exemplars.
5. For each subject area, descriptions were revised at a separate validation meeting by a group that included some St. Louis panelists and some new members.	
6. Descriptions were put in final form by NAGB staff and consultants.	
7. Final descriptions were adopted by NAGB.	7. Final cutpoints were adopted by NAGB. (In mathematics, the cutpoints based on judgments by St. Louis panelists were lowered by one standard error of measurement.)

To examine the adequacy of the process used to develop descriptions in reading, the Panel commissioned an evaluation study by reading experts, which included interviews and surveys of participants and observations of both the initial meeting in St. Louis and the followup validation meeting in San Diego.[1] The Panel also asked groups of content experts in both reading and mathematics to review the different versions of the descriptions. The results of the content-expert analyses are summarized briefly in this chapter, but are treated at greater length in chapter 4 as part of the external validity evaluation.

Evaluation of the Process for Setting Cutscores

As described in chapter 2, the achievement-level cutpoints were set using a modification of the widely known Angoff standard-setting procedure. The Angoff procedure makes strong assumptions about the ability of judges to conceptualize the performance of students at the borderline of each level to such a degree of specificity that they can estimate "p-values" for each separate item for these hypothetical examinees. For example, to make item ratings for the fourth-grade proficient cutpoint in mathematics, judges must hold in mind the definition of what fourth-grade proficient students should be able to do as stipulated in the description for that level, and they must envision a "borderline student" fitting this description (which means a student who just barely meets the standard for the level). Then for each item, judges must think about what skills or abilities are measured and about what features of the item format will make the item difficult or easy for fourth-grade examinees in the proficient category (however, judges are seldom given training in this aspect of the judging process). Finally, judges must translate these understandings into an estimated p-value for each item. P-value is a measurement term referring to the proportion of examinees who get an item right. In the Angoff methodology, p-values are estimates provided by the judges of the proportion of borderline students who will answer the item correctly.

As shown in figure 3.1, panelists went through three rounds of item ratings, changing their ratings, if they wished, in response to feedback that was provided at each stage. Final cutpoints on the NAEP scale, which become the minimum scores necessary to attain each of the levels, were then derived statistically by a procedure that translated p-value estimates for all the judges on all of the items into specific values on the NAEP scale. The statistical procedure is conceptually similar to averaging estimated p-values.

The level-setting process was also seen as a means for arriving at national consensus standards—with respect to both the descriptions and cutpoints—based on representative groups of educators and noneducators. The purpose of the process evaluation was to determine whether underlying assumptions held true in the actual development of the descriptions and in the implementation of the Angoff procedure. For example, the statistical procedure used to obtain cutscores can be used item-by-item to obtain a judge's intended cutscore. If individual judges are able to hold a consistent view of a hypothetical borderline student, then the cutscore implied by each item judgment should be about the same. Do judges' ratings in fact exhibit this

[1] Panel observers were also present at the initial level-setting meeting in mathematics as part of the ongoing evaluation of the Trial State Assessment. However, because the special study of the achievement levels had not been requested at that time, mathematics experts did not join the study team until the time of the validation meeting in mathematics.

kind of self-reliability? Further, if the process leads to consensus decisions, then we would expect to see more agreement in judges' ratings at the end of the process than during the first round of item ratings. Was such evidence of consensus obtained?

The Panel's investigation of these internal validity issues involved several independent studies and reanalyses of existing data. During the standard-setting meetings for reading and mathematics, NAGB's contractor kept detailed records of the judgments made by each judge on each item for each of the three rounds of the Angoff procedure. These data were made available to the Panel and were used to study systematic effects of item format, item content, and group membership on item judgments, and changes in judges' ratings from Round 1 to Round 3. Where pertinent, findings from the reading process evaluation were integrated with analyses of the judgment data. In addition, the Panel conducted two experimental studies to test the effects of manipulating key aspects of the judgmental process.

Organization of the Chapter

This chapter is organized chronologically to consider first the development of descriptions, then the setting of cutscores, and finally revisions of the descriptions and cutscores. This is followed by four sections that address these four research questions:

1. Did the process of developing descriptions and making item judgments in reading reflect the intended relationships among the Reading Framework, reading assessment, descriptions, and cutscores?

2. In reading and in mathematics, did the judgment process produce judgments that are internally consistent and coherent?

3. Is the judgment process affected by making three ratings at once or by making item-by-item ratings?

4. Did the process facilitate consensus?

In the final sections of the chapter, the Panel presents a critique of the Angoff method, an evaluation of the decisions to adjust or not adjust recommended cutpoints, and a discussion of the possible effects on item judgments of changing the achievement-level descriptions.

Relationships Among the NAEP Framework, Achievement-Level Descriptions, and Item Judgments in Reading

A key set of evaluation issues pertains to the relationship between the NAEP frameworks and the processes used for developing descriptions and for rating items.

Role of the Framework

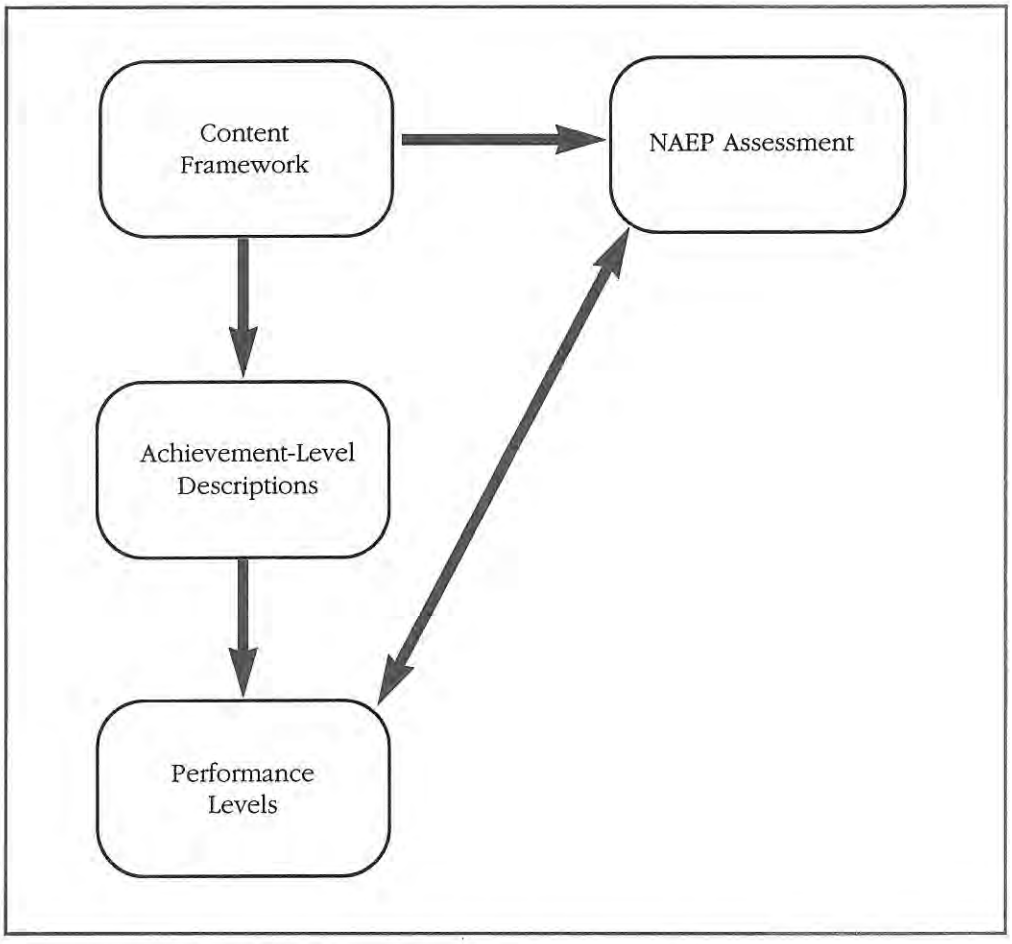
In order to evaluate the adequacy of the achievement-level-setting process, as it was implemented by NAGB and ACT, it is important to understand the assumed relationships between a subject-area framework (developed from a consensus process involving professional educators) and other elements of the assessment. First, given that the stated purpose of the framework is to guide the development of items for the assessment, there should be a close correspondence between the framework and the assessment items. Second, given that NAGB intends the achievement levels and the framework to remain stable over several administrations (while the actual items will change with each administration), the framework must serve as the pivotal link between the assessment and the achievement levels—both as they are described narratively and as they are operationalized into item-level judgments and ultimately cutscores. Ideally, this would imply that the achievement-level descriptions are crafted concurrently with the framework (as has been done for frameworks coming on line in 1994); but at the minimum, it requires that participants in the level-setting process be familiar with the framework and use it as the primary basis for developing the descriptions. The descriptions, in turn, should play a primary role during the item-rating process in helping judges identify the hypothetical groups that exist at each cutpoint.

The expected relationships among the content framework, the NAEP assessment, achievement-level descriptions, and performance levels are modeled in figure 3.2. As shown by the heavy dark lines, the framework determines both the assessment content and the achievement-level descriptions, while the latter, in turn, form the basis for the performance levels. A strong reciprocal relationship is also shown between the performance levels (cutscores) and the NAEP assessment, each influencing the other. *In the Panel's judgment, the integrity of NAEP interpretations based on achievement levels depends on the fidelity of the relationships depicted in the figure.*

From the evaluation study of the reading process, several critical problems were identified that violate the assumed relationships and therefore bear on the validity of the final descriptions and levels.[2] First, most participants were unfamiliar with the Reading Framework and relied instead on personal definitions of reading and personal experience to describe achievement levels. Second, the initial descriptions were influenced more by assessment items than by the framework. Third, many participants

[2] D. Pearson and L. DeStefano, "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report One: A Commentary on the Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993).

Figure 3.2. ***Model of relationships among components of the proposed achievement-level-setting process***



SOURCE: D. Pearson and L. DeStefano, "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report One: A Commentary on the Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993).

reported using referent groups from their own experience rather than using the descriptions to make item judgments. Findings that led to these conclusions are summarized briefly in the sections that follow.

Influence of the Framework on Descriptions

The Reading Framework for the 1992 NAEP represented major changes from previous assessments including a constructivist view of what it means to be a good reader, use of authentic texts, variation in reading situations, assessment of cognitive strategies, and an increased emphasis on open-ended questions. In contrast, when participants

who set the achievement levels were asked in a mail survey to describe a good reader 2 weeks after the level-setting session, they gave a great variety of answers; and most reflected a discrete subskills definition of reading out of keeping with the framework definition. Examples of descriptions of good readers written by participants are in table 3.1.

Table 3.1. ***Descriptions of good readers from participants after the St. Louis level-setting session***

Good decoding skills, can read the words, use other words to figure out unknown words. Can comprehend and then be able to answer questions about what was read.
One who can read with fluency and understanding without stumbling over words or passages typical of the grade level represented.
One who frequently reads with comprehension—has a wide vocabulary.
A good reader is one that is fluent. After reading a passage, they can then comprehend and apply information digested.
A good reader is able to comprehend the material, that is to decode and apply the reading to the content at the evaluation level (in accordance with Bloom's taxonomy).
Understands most vocabulary either on sight or by context. Able to relate past experiences with materials read.
I wish that I'd copied the definition we worked on in our meeting, but as I didn't...reading is the process of making meaning from print, of decoding words, interpreting the meaning of text through interplay with one's prior knowledge and experience.

SOURCE: D. Pearson and L. DeStefano, "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report One: A Commentary on the Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 24.

Although the NAEP framework definition of what it means to be a good reader includes fluency and comprehension, it also includes additional characteristics such as the following, which were mentioned by only a few participants or none at all: possessing positive habits and attitudes about reading; using prior knowledge to understand; being able to extend, elaborate, and critically judge the meaning of what is read; using a variety of effective strategies to aid understanding and to plan, manage, and check the progress of one's own reading; and being able to read a wide variety of texts and to read for different purposes. *Participants' lack of familiarity with the Reading Framework probably affected what they were able to hold in mind when*

making item judgments and most certainly explains why the achievement-level descriptions developed at the initial meeting had to be revised subsequently to be brought in line with the framework.

Influence of Assessment Items on Descriptions

Before they began their task of developing the achievement-level descriptions in reading, participants were given an overview of the Reading Framework and told about its definition of a good reader, which incorporates a combination of reading stances and reading situations. Observers of the process noted that most working groups subsequently picked up the framework and attempted to use it to generate descriptions. In particular, groups referred to the situations-by-stances matrix. However, participants appeared to be uncertain about how to translate elements in the matrix into descriptions of specific levels and "eventually, participants responded to time pressures by limiting discussion and placing more emphasis on generating lists of descriptors using brainstorming and other techniques... ."[3] During this process, participants were encouraged by facilitators to remember test content by thinking back to the portion of the 1992 assessment they had taken an hour earlier, and they were frequently admonished that "if it is not on the test, it can't be in the descriptors."[4] At the end of the process, participants also referred to copies of specific assessment items to revise the descriptions. *Participants' lack of an internalized and firm understanding of the framework, coupled with the instructions to attend to specific assessment content, most likely explain why the St. Louis descriptions were tied more closely to the 1992 assessment items than to the framework.* (An example of the three versions of the descriptions—for grade 12 basic performance—is shown later in figure 3.6.) The evaluators of the process judged this emphasis on specific assessment content to be problematic because the 1992 reading items did not, in fact, have the desired relation to the framework.

Influence of Personal Experience on Item Judgments

In theory, assessment items are written to reflect the assessment framework. Therefore, it should make little difference whether participants developed descriptors by focusing on test items or the framework. However, because of the short timelines for developing assessment items in response to the 1992 Reading Framework, *content experts found a number of misalignments between the framework and actual NAEP items, especially with respect to measuring different stances and different reading situations.*[5] This poses a serious dilemma. If the descriptions are based on the framework, then NAEP interpretations based on the descriptions will be misleading

[3] Ibid., 28.

[4] Ibid., 25.

[5] B. Bruce, J. Osborn, and M. Commeyras, "The Content and Curricular Validity of the 1992 National Assessment of Educational Progress in Reading" (Stanford, CA: The National Academy of Education, forthcoming).

because they refer to some abilities not even assessed. If, as occurred, the descriptions are based on the actual sample of items in the baseline assessment, then they may not be appropriate for describing changes over time. The decision to revise the descriptions to align more with the Reading Framework is discussed in a subsequent section of this chapter.

In figure 3.2, the expected model for developing the achievement levels, it is assumed in the last step that the descriptions based on the framework are used to make item judgments. In the reading process study, interview and written survey data were used to analyze how participants went about rating items. Only one third of the 36 respondents described a process that included referring to the descriptions. A disproportionate number of these respondents (9 of the 12) were representatives of the general public or nonteacher educators. As explained by one such respondent, "I used the descriptors to decide on a rating. Since I was a 'noneducator' participant, I relied heavily on the descriptors."[6] In contrast, only 3 of 21 teachers interviewed mentioned descriptions as part of their rating process. The majority of teachers explained using their own experience with students to make judgments about items. Examples of teacher responses of this type are seen in table 3.2.

Table 3.2. ***The bases for teacher judgments about items at the St. Louis level-setting session***

I rated each one according to what the students in my class would be able to do.
I applied increasing amounts of evidence—own experiences, others' discussion, etc. If I was in doubt, I went with my own experience and gut-level feeling.
I used my personal experiences—I compared that test to the tests we use in [home state].
I thought of students in my class and how many would be able to answer the questions correctly.

SOURCE: D. Pearson and L. DeStefano, "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report One: A Commentary on the Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 33.

These responses suggest that many participants were not making systematic judgments based on specific features of the descriptions. They may also suggest, ironically, that participants were using normative performance standards to make their item ratings rather than judging substantively what students should *be able to do.* The last quotation, for example, where a teacher is making judgments based on "how many" of his or her students could do the item, is clearly normative. Alternatively, one teacher (in response to a different question) clarified that he (she) thought of students who would fit the description or the level and then estimated how that student or group of students would do on the item. The second teacher's use of familiar students as a referent group does not imply that the standards are being set normatively.

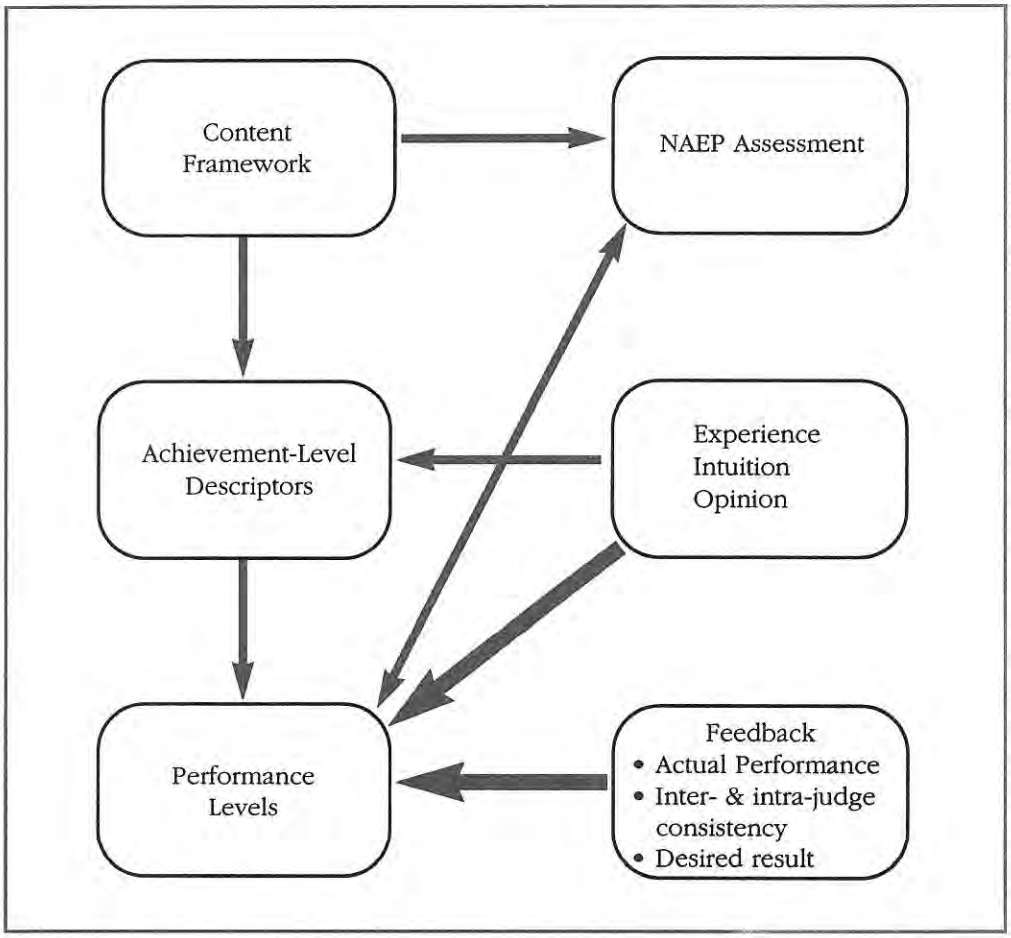
6 Pearson and DeStefano, op. cit., 32.

Unfortunately, in most cases it is not at all clear how teachers were using information from students they knew to arrive at p-value estimates.

A Model of the Observed Achievement-Level-Setting Process

On the basis of these findings, *the evaluators of the reading level-setting process concluded that the idealized model and expected relationships had not been realized in practice. Participants simply did not know the Reading Framework well enough for it to have the desired influence on the development of the reading descriptors or item judgments.* The evaluators developed an alternative model, presented in figure 3.3, to illustrate what factors influenced development of descriptors and item judgments in

Figure 3.3. *Model of relationships among components of the observed achievement-level-setting process*



SOURCE: D. Pearson and L. DeStefano, "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report One: A Commentary on the Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 44.

actual practice. In the model of the observed process, the expected relationships between the NAEP content framework, the NAEP assessment, the achievement-level descriptions, and the cutscores are still shown, but their influence is much weaker than intended. Two other factors, which were not intended to be a part of the achievement-level-setting process, have been added to the model. These added, unintended factors have much greater impact on the level descriptions and cutscores: (1) judges' personal experience and opinion and (2) feedback provided to judges during the three rounds of the Angoff procedure. The kind of feedback provided to judges and its effect on cutscores are analyzed later in this chapter.

The Consistency and Coherence of Item Judgments

What should item data look like if judges are making item judgments in an internally consistent manner? Judges are *not* expected to estimate p-values for borderline proficient or borderline basic students that are the same as real-data p-values. Judges are expected to envision performances that *should* be; therefore, their judgments do not necessarily have to conform to normative data. However, their judgments should be internally consistent as a reflection of their expectations for borderline performance. We would expect similar ratings for items measuring the same content, but judges might have good reasons to vary their ratings over different content dimensions. For example, they might make systematically higher ratings (compared to empirical scale values) for items on a particular topic if they believed that topic had not been taught but should be taught. However, item ratings (and the implied cutscores) should not vary in response to irrelevant item features such as variations in item format.

The Panel commissioned a series of five "studies" or reanalyses of data collected during the level-setting process to examine the internal consistency of judges' ratings across different types of items.[7] The first three of these studies addressed the effects of item-type differences: right-wrong dichotomous versus extended-response, four-point-rubric-scored items; multiple-choice versus short-answer items; and easy versus hard items. The effects of item content and cognitive processes were addressed in two additional studies. For both reading and mathematics, the data were the original data gathered by ACT as the actual 1992 NAEP achievement levels were being set. For the purposes of these five studies, only the results based on Round 3 data are reported, which means that the judges had already received feedback about item difficulty and internal consistency and had finished their level-setting task.

Impact on Cutscores of Right-Wrong Versus Extended-Response Items

The first comparison tests the consistency of judges' ratings when both the type of item and the judges' instructions were different. The modified Angoff approach was used to set cutpoints for right-wrong items, which include both multiple-choice and

[7] D.H. McLaughlin, "Validity of the 1992 NAEP Achievement-Level-Setting Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993).

short-answer type questions. The Angoff procedure is not directly applicable for rating items scored on a 1-to-4-point scale. Therefore, for extended-response items, judges were given several dozen examples of actual student responses and were asked to pick papers that represented borderline basic, borderline proficient, and borderline advanced performance.

The data for these comparisons are shown in table 3.3M for mathematics and 3.3R for reading. Results have been averaged over items and judges and translated into an estimated cutscore on the NAEP scale. This was done separately for the two types of items being compared. What can be seen in both tables is that *ratings of extended-response items lead to substantially higher cutpoints.* This pattern is true for all achievement levels, for all grades, and for both reading and mathematics. In practical terms, these differences are huge and argue against the credibility of one or both sets of judgments. *For example, the cutpoint set for fourth-grade basic in mathematics based on right-wrong questions was 210.4 and would identify 61 percent of fourth graders as basic or above. The cutpoint set for fourth-grade basic in mathematics based on extended-response items was 266.5, which is higher than the eighth-grade basic cutoff based on right-wrong items, and would find only 6 percent of fourth graders to be basic or above.* Keep in mind that these differences in implied standards are not due to the greater difficulty of extended open-ended questions compared to right-wrong questions covering the same material. The statistical procedures used to estimate intended cutscores already take the relative empirical difficulty of the items into account.

Some relative difference in the standards set by item type is expected if, for example, panelists want to emphasize skills not currently taught in classrooms. However, the obtained differences are so large that it is implausible to explain them on these grounds. *Instead, the extreme differences between cutpoints based on right-wrong items and extended-response items argue strongly that judges are unable to maintain a consistent view of borderline performance for each of the levels.* The discrepancies in judgments may be due either to differences in judges' perceptions of performance on the two item types or to the different methodologies used to collect the data. ACT attempted to address this issue by conducting a "Round 4" study with only the extended-response items. By instructing panelists to consider scores of 2, 3, or 4 as correct and then estimate p-values, cutpoints were brought nearly in line with the right-wrong items in most cases. However, these Round 4 results were not used in the final calculations of the achievement levels. Although the Round 4 study was by no means conclusive, it strongly suggests that panelists were not setting higher levels for exemplar papers based on some conceptual or principled choice; otherwise it would not be so easy to change the cutscores by changing the methodology.

It is the Panel's view that the large discrepancies caused by differences in item types reflect on the nature of the judgment task itself. Differences cannot simply be corrected by statistical adjustment or by manipulation of instructions to achieve consistency. Procedural or technical remedies should not be applied when the "errors" are erratic and inexplicable. For example, it cannot be assumed that the Angoff, right-wrong results are correct; therefore, the extended-response results should not be adjusted to correspond to them. Greater understanding should be sought of how judges conceptualize basic, proficient, and advanced performance when identifying boundary exemplars, how these conceptions are influenced by the sample of papers available, and what assumptions judges are making in generalizing from a single paper to a student's total proficiency.

Table 3.3M. ***Mathematics achievement-level cutpoints, based separately on dichotomously scored items (Angoff procedure)and extended-response items (Boundary Exemplars procedure)***

	Grade 4		Grade 8		Grade 12	
	Scale Score	Percent At Or Above	Scale Score	Percent At Or Above	Scale Score	Percent At Or Above
Basic						
Dichotomous Items	210.4	61	255.6	63	288.9	61
Extended-Response Items	266.5	6	302.7	17	344.3	9
Proficient						
Dichotomous Items	250.0	16	297.5	21	333.9	16
Extended-Response Items	304.4	0.2	345.5	1	374.0	1
Advanced						
Dichotomous Items	282.7	2	333.3	3	366.2	2
Extended-Response Items	330.8	0.01	376.8	0.03	388.0	0.2

Table 3.3R. ***Reading achievement-level cutpoints, based separately on dichotomously scored items (Angoff procedure)and extended-response items (Boundary Exemplars procedure)***

	Grade 4		Grade 8		Grade 12	
	Scale Score	Percent At Or Above	Scale Score	Percent At Or Above	Scale Score	Percent At Or Above
Basic						
Dichotomous Items	190.6	78	232.5	78	249.9	88
Extended-Response Items	281.1	3	290.1	20	329.1	12
Proficient						
Dichotomous Items	229.6	39	272.1	38	293.6	48
Extended-Response Items	317.4	0.1	336.4	1	362.6	1
Advanced						
Dichotomous Items	259.9	12	311.0	7	336.8	7
Extended-Response Items	356.2	0.00	388.8	0.00	393.6	0.01

SOURCE: D.H. McLaughlin, "Validity of the 1992 NAEP Achievement-Level-Setting Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 88.

For example, one possible explanation for the systematically high standards set from the extended-response items is the nonrepresentative sample of exemplar papers provided to the judges. The papers were nonrepresentative of the actual distribution of student performance in two ways. First, to provide sufficient examples of potentially advanced performance, papers receiving high scores, which were rare in the assessment, were over-sampled.[8] Second, nonresponsive and off-target papers, although frequent in the assessment, were omitted because they would be uninformative.[9] Thus, judges were looking at a systematically better sample of student papers than for students in general. Of course, if this were the problem, it could be "fixed" superficially in the next standard-setting effort by using a representative sample of papers. However, this explanation, if true, would imply a more fundamental problem—namely that judges are making normative and comparative judgments, strongly influenced by the sample of papers they see, rather than making judgments based on the descriptions and their conceptions of what students should be able to do.

Another possible explanation for the huge differences between extended-response and right-wrong cutpoints is that judges rely on the same assumption made by the statistical methodology, namely that students consistently produce papers like the one observed with 100 percent reliability. Similar problems have been encountered with the effort to set cutpoints for the achievement levels in writing, suggesting that these effects are systematic and should be investigated not only by NAGB and its contractors but also by the larger measurement community. This issue also raises questions about the effect on conceptions of proficiency of making item-by-item judgments under the Angoff method. (What assumptions are judges making implicitly about the reliability and intercorrelation of items?) In a later study, we consider how students are evaluated when judges have access to more extended and integrated evidence of student work.

Impact on Cutscores of Multiple-Choice versus Short-Answer Items

Two other comparisons allowed an evaluation of the effects of item type without the confounding effect of changes in judgmental method. Recommended cutpoints were compared for multiple-choice versus short-answer questions and for easy versus hard questions. All of these types of items are scored right-wrong and were rated using the Angoff procedure. Resulting cutpoints for multiple-choice versus short-answer items are presented for mathematics and reading in tables 3.4M and 3.4R respectively. As can be seen, judgments based on short-answer items almost always lead to higher cutpoints than judgments based on multiple-choice items. These differences are not, however, as extreme as those observed between cutpoints based on dichotomous versus extended-response items.

One criticism of the process for setting achievement levels in 1990 was that judges had not been given appropriate instructions about how to take guessing into account. This

[8] Specifically, extended responses that contained any substantial content relevant to an item were scored 1, 2, 3, or 4; and a roughly equal number of papers with each score were shown to each judge.

[9] For some items, more than 25 percent of the responses were either blank or completely off target.

Table 3.4M. ***Mathematics achievement-level cutpoints, based separately on multiple-choice items and short-answer open-ended items***

	Grade 4		Grade 8		Grade 12	
	Scale Score	Percent At Or Above	Scale Score	Percent At Or Above	Scale Score	Percent At Or Above
Basic						
Multiple-Choice Items	190.8	80	246.3	71	278.1	72
Short-Answer Items	218.3	52	257.4	61	307.3	41
Proficient						
Multiple-Choice Items	246.7	19	294.6	24	324.3	24
Short-Answer Items	246.9	19	296.1	23	343.6	9
Advanced						
Multiple-Choice Items	282.8	2	330.3	4	358.4	3
Short-Answer Items	275.8	3	334.7	3	369.2	1

Table 3.4R. ***Reading achievement-level cutpoints, based separately on multiple-choice items and short-answer open-ended items***

	Grade 4		Grade 8		Grade 12	
	Scale Score	Percent At Or Above	Scale Score	Percent At Or Above	Scale Score	Percent At Or Above
Basic						
Multiple-Choice Items	175.9	87	184.6	97	219.7	98
Short-Answer Items	199.7	70	240.0	72	258.7	83
Proficient						
Multiple-Choice Items	222.7	47	259.1	53	283.1	61
Short-Answer Items	234.5	33	281.2	29	296.4	45
Advanced						
Multiple-Choice Items	255.4	15	296.2	15	327.8	13
Short-Answer Items	263.4	10	322.2	3	337.9	7

SOURCE: D.H. McLaughlin, "Validity of the 1992 NAEP Achievement-Level-Setting Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 94-95.

problem recurred with the 1992 achievement levels. No explicit instructions were given to the judges about whether to allow for guessing when estimating the proportion of each borderline group that would be able to get the item right. In addition, judges may not recognize a phenomenon that is well known in the psychometric literature. Supplying answer choices typically makes a question much easier (apart from guessing) because it structures the problem and permits the student to pick the right answer, to catch mistakes, and so forth. (When items that appear to be similar in content have very different empirical difficulties, translation into scale values will result in very different cutpoints, unless judges are able to take these format effects into account.)

As might be expected, differences between short-answer and multiple-choice item types have the most pronounced effect on the difficulty of items for low-scoring students because even students who have no understanding of the problem have a reasonable chance of getting a multiple-choice item right by guessing or random marking. As a consequence, *there were sharp differences in cutpoints set for the basic level for all three grades and across subject areas. Multiple-choice items produced significantly lower cutpoints than did short-answer items, resulting in up to 30 percent more students being classified as basic or above.* For example, in fourth-grade mathematics, the two cutpoints would mean the difference between 80 percent versus 52 percent of fourth graders being at the basic level or above. *These large differences again raise questions about the ability of judges to make item ratings in accord with hypothetical conceptions of student proficiency.* Conceptual inconsistencies are especially problematic given the likelihood that the assessment will change from year to year to include more extended-response and short-answer type questions.

Impact on Cutscores of Easy Versus Hard Items

The data in tables 3.5M and 3.5R reflect a third comparison of the effect of item type on Angoff judgments. Recommended cutpoints were compared based on easy versus hard items (with the effects of multiple-choice versus short-answer questions held constant). As can be seen, item difficulty (easy versus hard) had a large and significant effect on judges' cutpoints. This means that after taking item difficulty into account (which is done by the cutpoint estimation methodology), the judges were still pointing to very different places on the NAEP scale for their intended standards based on easy and hard items. While judges estimated higher p-values for easy items than for hard items (as evidenced by correlations between judges' ratings and real-data p-values), judges failed to adjust sufficiently for differences in item difficulty. *Panelists tended to underestimate performance on easy items and overestimate it for hard items. Therefore, rather than setting a consistent expectation for each level, panelists generated very different cutpoints from the easy and hard items.* After adjusting for item difficulty statistically, panelists were still asking for relatively easy standards on easy items and for differentially difficult standards on hard items. In mathematics, differences in the implied cutpoints could change the percentage of fourth-grade students said to be advanced from 2 percent to 10 percent, or change the percentage of proficient students from 16 percent to 46 percent. These findings were consistent across grades and were of even greater magnitude in reading. Moreover, they were consistent with data reported in the GAO study for 1990 mathematics results, which found that examinees consistently exceeded item-judgment expectations for "easy" and

Table 3.5M. ***Mathematics achievement-level cutpoints, based separately on easy items (difficulty < median) and hard items (difficulty > median), excluding extended-response items***

	Grade 4		Grade 8		Grade 12	
	Scale Score	Percent At Or Above	Scale Score	Percent At Or Above	Scale Score	Percent At Or Above
Basic						
Easy Items	187.4	82	240.2	76	273.4	76
Difficult Items	205.8	66	270.8	47	302.8	46
Proficient						
Easy Items	223.2	46	271.0	47	306.8	42
Difficult Items	250.2	16	306.6	14	338.3	12
Advanced						
Easy Items	258.2	10	301.6	18	341.4	10
Difficult Items	281.0	2	339.2	2	365.5	2

Table 3.5R. ***Reading achievement-level cutpoints, based separately on easy items (difficulty < median) and hard items (difficulty > median), excluding extended-response items***

	Grade 4		Grade 8		Grade 12	
	Scale Score	Percent At Or Above	Scale Score	Percent At Or Above	Scale Score	Percent At Or Above
Basic						
Easy Items	179.6	85	202.8	93	228.2	96
Difficult Items	195.6	74	231.4	79	246.2	90
Proficient						
Easy Items	213.5	57	254.0	59	273.8	70
Difficult Items	240.6	27	285.0	25	304.7	35
Advanced						
Easy Items	242.8	25	286.8	23	314.1	25
Difficult Items	266.8	8	320.8	3	342.5	5

SOURCE: D.H. McLaughlin, "Validity of the 1992 NAEP Achievement-Level-Setting Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 96-97.

"moderately easy" items and fell short of expectations on the most difficult items. The consistency of findings across years and subject areas suggests that differences by item type are an artifact of the judgmental method or reflect biases in judges' estimates of item difficulty.

Ultimately, the standard of performance expected of students should not be an artifact of the judgmental process. Different standards (cutpoints) following from different expectations by subcontent area would be defensible; differences in standards that appear to be caused by arbitrary features of the standard-setting procedures themselves are unacceptable. Simply averaging the results across types of items to arrive at the final cutpoints does not alleviate the inherent inconsistencies. Even more important, differences between item types such as those observed threaten the measurement of progress over time because any change in the mixture of these items in the assessment will invalidate trend data.

Impact on Cutscores of Content- and Cognitive-Process Subscales

The Panel expected that there could be substantive explanations for systematic differences in item ratings. For example, suppose that measurement and probability are taught less frequently than numerical operations. If judges set cutpoints to reflect what students *should* know, it would be reasonable for cutpoints based on unfamiliar subjects to be set higher in relation to the empirical scale than those based on the familiar content strand. Recommended cutpoints are reported in table 3.6M separately for the five mathematics subscales (1. numerical operations, 2. measurement, 3. geometry, 4. probability, and 5. algebra) and in table 3.6R for the three reading subscales (1. reading for literary experience, 2. reading for information, and 3. reading to perform a task). Surprisingly, *substantive differences did* not *lead to large differences in cutscores,* with the exception of the proficient and advanced levels for eighth-grade mathematics, where higher cutscores were set for measurement and for data analysis, statistics, and probability than for other scales.

Similar analyses were done in a fifth study contrasting cutpoints set to represent different cognitive processes. These process categories are identified as part of the respective content frameworks and are used in developing assessment items to ensure a proper balance among items tapping higher- and lower-level thinking skills. In mathematics, the processes are procedural knowledge, conceptual understanding, and problem solving; in reading, they are forming an initial understanding, developing an interpretation, reflecting on the text, and demonstrating a critical stance. Again *the results did not suggest any systematic variation in recommended cutpoints due to substantive dimensions of the assessment.*

Like the findings for the content dimensions of the tests, the finding of no difference in cutpoints by cognitive category is surprising because it suggests that judges are implicitly "content" with the current relationship of higher-order and lower-order thinking skills as represented by the empirical relations of the items. If they had been discontent, wishing to see greater gains in the future on items demanding higher levels of reasoning (such as problem solving or demonstrating a critical stance), they would have set higher cutpoints on the basis of these items. Another possible

Table 3.6M. ***Mathematics achievement-level cutpoints, separately by subscale***

Subscale	Grade 4					Grade 8					Grade 12				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Including All Items															
No. of Items	63	29	27	20	15	58	32	36	28	29	42	28	32	29	47
Basic	209	214	213	221	216	255	256	252	265	256	283	300	294	287	299
Proficient	250	256	248	250	253	298	314	288	315	296	327	337	336	334	341
Advanced	284	288	281	274	284	335	358	322	350	330	359	364	365	374	370
Excluding Extended-Response Items															
No. of Items	61	29	26	19	14	56	31	35	27	28	42	28	31	28	43
Basic	203	214	213	214	207	254	252	250	258	252	283	300	292	286	293
Proficient	247	256	247	246	247	294	306	285	308	290	327	337	336	333	335
Advanced	282	288	277	272	279	324	351	319	345	325	359	364	364	372	364

Table 3.6R. ***Reading achievement-level cutpoints, separately by subscale***

Subscale	Grade 4			Grade 8			Grade 12		
	1	2	3	1	2	3	1	2	3
Including All Items									
No. of Items	56	44	0	35	52	48	32	62	51
Basic	212.2	201.6	—	235.5	241.6	245.3	259.9	268.8	261.8
Proficient	244.2	236.4	—	280.9	278.5	286.6	302.4	305.2	304.9
Advanced	278.2	267.9	—	330.6	328.2	331.0	356.1	343.9	350.6
Excluding Extended-Response Items									
No. of Items	49	38	0	32	46	43	29	53	46
Basic	195.1	187.2	—	227.1	232.8	237.3	246.4	256.4	249.6
Proficient	230.9	224.5	—	267.1	268.7	279.3	290.5	292.6	294.1
Advanced	261.2	254.3	—	304.7	306.2	318.7	339.2	329.9	337.3

SOURCE: D.H. McLaughlin, "Validity of the 1992 NAEP Achievement-Level-Setting Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 92.

interpretation of these results is that judges are not differentiating their expectations for the items by content at all and are merely responding to item difficulty as best they are able.

Impact on Cutscores of Feedback Provided to Judges

Recall that the preceding studies were all based on reanalysis of judges' ratings collected during the actual 1992 level-setting meetings. To interpret these findings further, it is useful to recall as well what information was provided to judges as they were making their ratings. Logical inconsistencies in judges' recommended standards are especially troubling given the number of procedural aids that were built into the process to facilitate internal consistency. For example, in Round 2 of the Angoff procedure, judges were given item-difficulty information. As expected, correlations between judges' estimated p-values and actual item p-values increased from Round 1 to Round 2.

Judges' self-consistency data increased between Round 1 and Round 2. These data are shown in tables 3.7M and 3.7R. Specifically, the tables show the average correlations between a judge's estimated p-value for a given item at a given level and the item p-value implied by the judge's overall estimate of performance for that level. Although judges tended to rank-order items correctly, the previous analysis of easy and hard items shows that they were recommending substantially different cutpoints based on easy versus hard items. Figure 3.4 is provided to illustrate how judges could produce item ratings that correlate well with empirical item p-values, but nonetheless contain systematic biases in the standards implied by hard and easy items.

For Round 3, judges were given personalized information about the items they were rating the most inconsistently compared to their own implied standards. Again this information served to improve consistency correlations slightly from Round 2 to Round 3, but did not correct the large artifacts created by different types of items.

After each round, judges were also shown a summary of results across grades to increase awareness of the need for grade-to-grade coherence. (The phase-one attempt to set mathematics achievement levels in 1990 had been criticized because the 8th-grade basic achievement level was set, illogically, higher than the 12th-grade basic level.) For the 1992 effort, panelists were shown idealized graphs reflecting the desired relations between levels within grades and increasing trends across grades. Then, using this same format, they were shown the results of their own ratings translated into cutpoints and compared across grades. Examples of the graphs presented after each of three rounds for reading are shown in figure 3.5. According to observations of the process in reading, the presentation of these graphs was highly suggestive. "As their own ratings began to resemble this prototype, participants were congratulated by the facilitators, were pleased with themselves, and gained confidence in the process. After each round, participants eagerly awaited the presentation of these graphs as a measure of the quality of their ratings."[10]

[10] Pearson and DeStefano, op. cit., 34.

Table 3.7M. ***Average correlations between a panelist's actual percent-correct estimates in mathematics and the percent-correct estimates implied by the panelist's maximum likely cutpoint scale value, by rounds***

	Grade 4			Grade 8			Grade 12		
Round	1	2	3	1	2	3	1	2	3
Basic	.43	.70	.75	.57	.78	.82	.59	.82	.85
Proficient	.36	.65	.72	.55	.76	.79	.56	.74	.77
Advanced	.25	.49	.60	.48	.64	.70	.42	.56	.61

Table 3.7R. ***Average correlations between a panelist's actual percent-correct estimates in reading and the percent-correct estimates implied by the panelist's maximum likely cutpoint scale value, by rounds***

	Grade 4			Grade 8			Grade 12		
Round	1	2	3	1	2	3	1	2	3
Basic	.36	.62	.66	.35	.63	.65	.33	.56	.57
Proficient	.31	.63	.66	.38	.67	.70	.32	.60	.62
Advanced	.24	.57	.61	.30	.62	.70	.23	.50	.54

SOURCE: D.H. McLaughlin, "Validity of the 1992 NAEP Achievement-Level-Setting Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 103.

Figure 3.4. ***An illustration of a relationship between actual item p-values and judges' estimated p-values with judges selecting less extreme p-values***

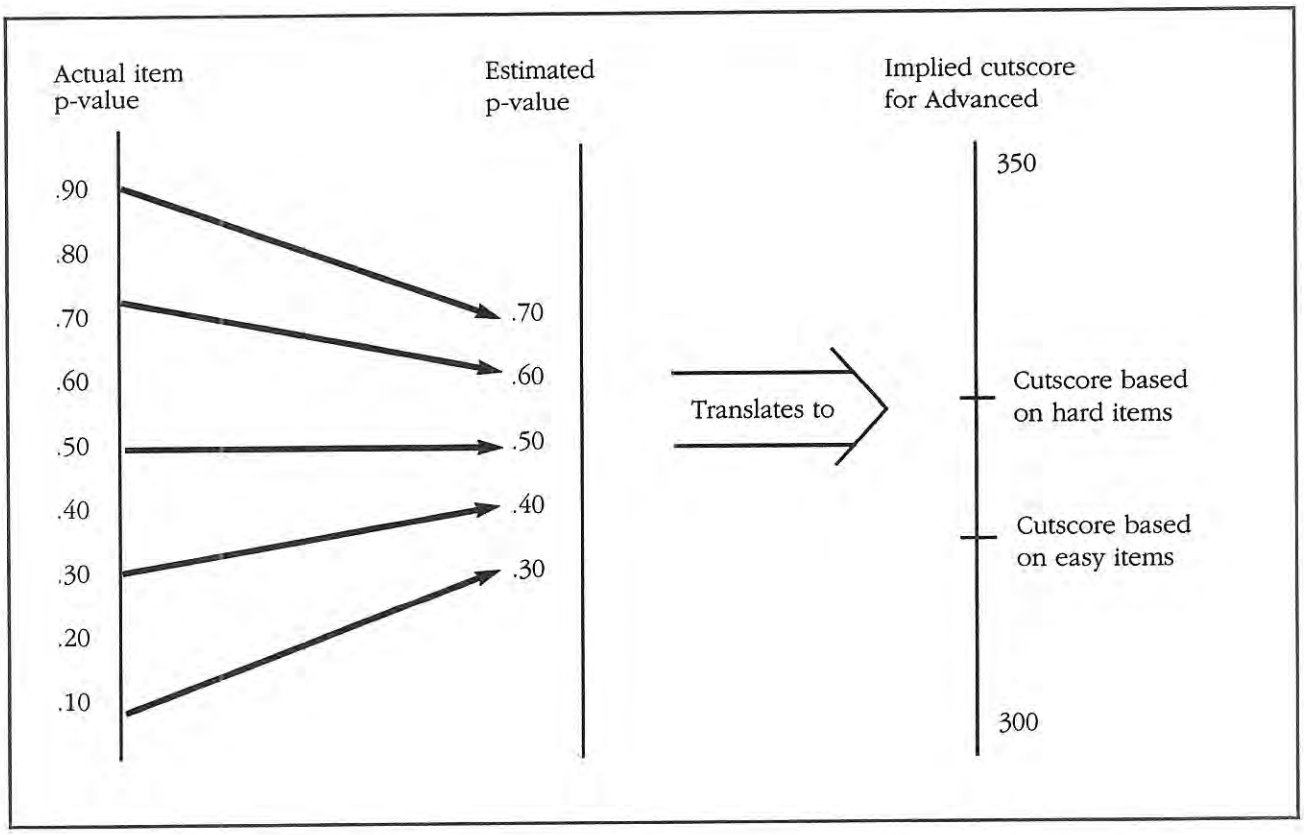
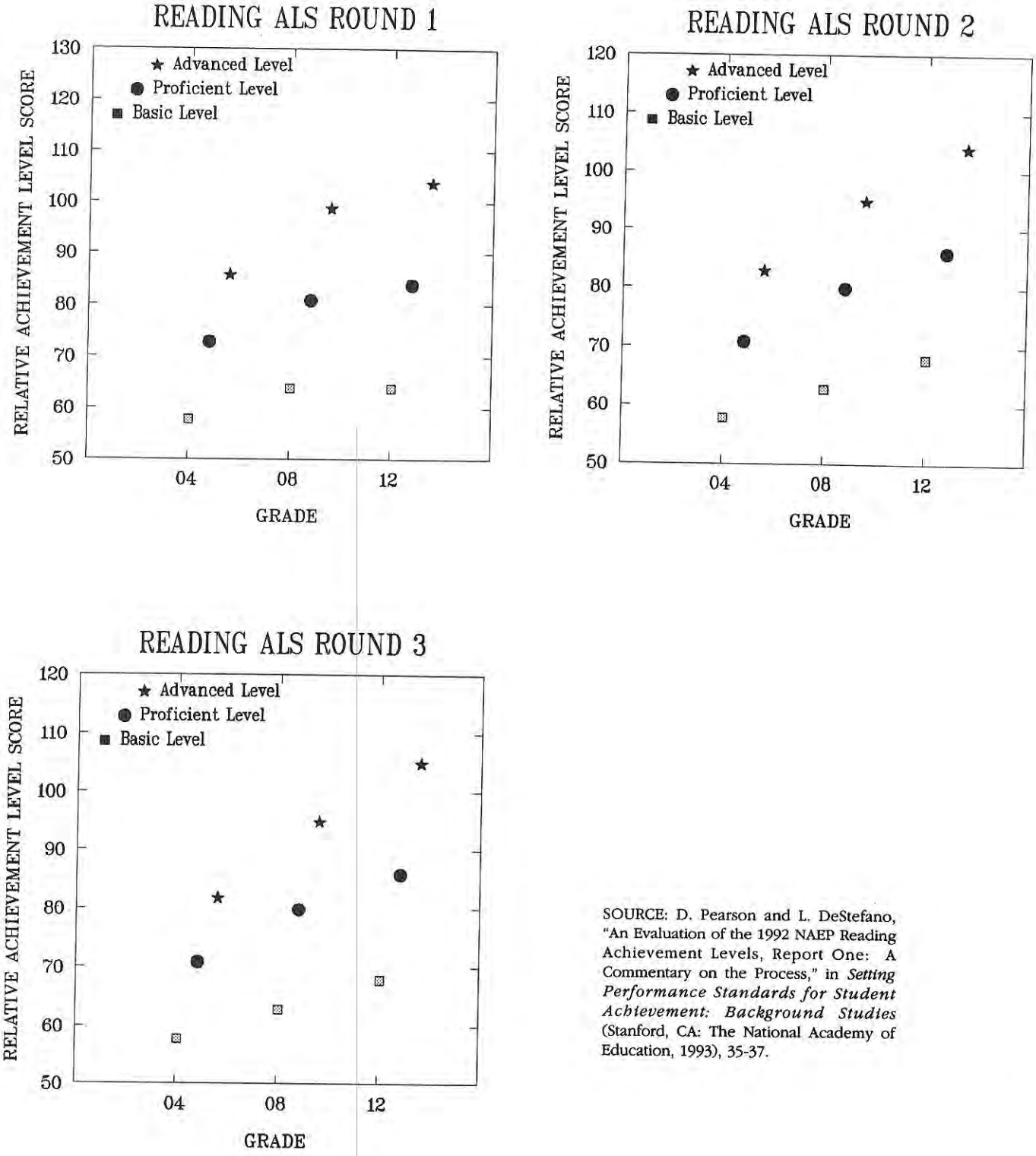


Figure 3.5. ***Cutpoints for three achievement levels across three grades in reading as reported to participants after each of the three rounds of ratings***



SOURCE: D. Pearson and L. DeStefano, "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report One: A Commentary on the Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 35-37.

The final levels in reading and mathematics follow a reasonable pattern across grades. However, given the procedures that were followed, it is not possible to conclude that the pattern emerged intrinsically from the judges' expectations about grade-level standards and growth from grade-to-grade. Rather, judges were conforming to a model that had been established as reasonable in previous evaluations. This is not to say that judges made wholesale changes in their ratings to conform or that facilitators were coercive. Participants were told repeatedly that they did not have to change their ratings in response to the different data sources they were shown as feedback. A more plausible conjecture about how the process worked is that judges were making cognitively difficult judgments and, as suggested by the data on judges' internal consistency, each judge experienced a great deal of uncertainty. Therefore, within a range of p-values that individual judges were contemplating, participants were probably willing to select values that were still consistent with their own beliefs but moved in the desired direction as implied by feedback.

Experimental Studies of Process Effects

In general, the Panel's reanalyses of rating-process data addressed questions about how procedures might affect judgments. Two additional questions could not be addressed with existing data; therefore small-scale experimental studies were undertaken. In the first of these studies, the effect of setting three cutscores at once was examined. In the second, whole-booklet judgments of student performance were compared to classifications of students based on item-by-item judgments.

Effects of Three Cutpoints on Item Ratings

One of the major modifications of the Angoff method for purposes of setting achievement levels was prompted by the need to establish lower boundary cutpoints for three different levels—basic, proficient, and advanced. Moreover, to set these three levels, the decision was made to have judges go through the item pool only once (per round) and estimate three different p-values per item corresponding to the expected performance of borderline students at each level. To what extent are judgments for each of the levels influenced by the need to set three cutpoints at once? Does this process create some systematic relationship among the three p-values? For example, do judges think about one estimated p-value and then choose the other two p-values in relation to it? Or do they conceptualize each level just as they would if they only had to set the standard for that level? In the experimental study, two randomly equivalent groups of judges were asked to set cutpoints for eighth-grade mathematics using the Angoff procedure.[11] In one group, the three points were estimated concurrently just as they were in the actual level-setting process. In the other group, all of the p-values were estimated for the basic cutpoint before going on to proficient, and so forth. Although two of three results were marginally statistically significant, the magnitude of effects did not argue strongly against the practice of making three judgments at once.

[11] D.H. McLaughlin, "Order of Angoff Ratings in Multiple Simultaneous Standards," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993).

Judging Whole Booklets Instead of Test Items

Another experimental study was conducted to examine the effects on cutscores of variations in verbal descriptions (which we discuss later in the chapter) and the effects of whole-booklet ratings.[12] It was hypothesized that allowing judges to see intact test booklets would provide more complete and integrated evidence of student performance and might lead to different conclusions about basic, proficient, and advanced levels of achievement than trying to create expectations by summing individual item probabilities. *When the same group of judges set cutpoints using the two different judgment methods, whole-booklet ratings led to a substantially higher cutpoint for the basic level and a substantially lower cutpoint for advanced (differing by 30 and 20 NAEP scale points respectively).* Although it is not possible to conclude definitively which cutpoints are the correct ones, this finding illustrates how the judgment task can affect conceptualization of each level, even using identical verbal descriptions. For example, a lower cutpoint for advanced is reasonable if students are expected to demonstrate advanced understandings but are not required to perform at that level with perfect reliability (i.e., some advanced items might be answered incorrectly). Differences in conceptualizations can in turn lead to quantitative differences in cutscores.

Adequacy of the Consensus Process

Two sources of data were available to permit an evaluation of the level-setting effort as a consensus process: item ratings collected from judges during the process and the reading observation study. Groups of panelists had been selected for each grade level to represent educators and noneducators, various regions of the country, and minority and majority groups. Of course, a consensus does not ensure the validity of standards; in particular, consensus does not ensure correspondence between the content of descriptions and empirical cutscores. But consensus is used to defend content choices and policy decisions. Were representatives of these diverse groups able to come together and agree on national performance standards for reading and mathematics?

Results for each of the three rounds of the Angoff procedure are reported in tables 3.8M, 3.8R, 3.9M, and 3.9R. Average cutscores for reading and mathematics are shown for each level in the first pair of tables. Standard deviations of judges' ratings are shown in the second pair of tables. It is apparent that, on average, recommended cutscores changed very little from the first round to the third and final round. Sometimes the cutscores went up a little, sometimes they went down a little, but overall the final recommended cutscores were very close to first-round judgments when judges were making their ratings independently. The largest shifts from Round 1 to Round 3 were in the fourth-grade advanced cutpoints, where the cutpoint was reduced by 9.5 points in mathematics and 14 points in reading; more typical were changes of 5 or 6 points from the first to the final round. The only consistent patterns were for 4th-grade standards to go down and 12th-grade standards to go up as the

[12] D.H. McLaughlin, "Rated Achievement Levels of Completed NAEP Mathematics Booklets," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993).

Table 3.8M. ***Mathematics achievement-level cutpoints, by rounds***

	Grade 4			Grade 8			Grade 12		
Round	1	2	3	1	2	3	1	2	3
Basic	215.8	207.4	212.8	248.3	250.4	256.4	285.3	288.8	292.4
Proficient	257.6	248.0	251.1	294.2	294.2	300.7	333.8	332.4	335.0
Advanced	292.9	280.5	283.4	329.5	329.3	337.1	366.6	363.4	365.9

Table 3.8R. ***Reading achievement-level cutpoints, by rounds***

	Grade 4			Grade 8			Grade 12		
Round	1	2	3	1	2	3	1	2	3
Basic	212.5	206.1	207.0	242.4	239.6	240.8	254.8	260.6	264.8
Proficient	247.2	241.0	240.5	285.3	280.6	282.9	296.5	302.6	304.6
Advanced	287.5	275.7	273.4	337.5	328.5	330.5	349.2	347.1	350.3

Table 3.9M. ***Mathematics achievement-level cutpoint standard deviations, by rounds***

	Grade 4			Grade 8			Grade 12		
Round	1	2	3	1	2	3	1	2	3
Basic	21.0	21.8	18.5	28.8	19.1	16.5	25.7	19.0	18.4
Proficient	16.9	16.9	14.7	20.7	17.5	16.6	14.0	14.1	13.9
Advanced	18.3	15.9	13.2	19.5	15.9	17.0	11.6	12.4	12.7

Table 3.9R. ***Reading achievement-level cutpoint standard deviations, by rounds***

	Grade 4			Grade 8			Grade 12		
Round	1	2	3	1	2	3	1	2	3
Basic	19.6	12.9	9.4	15.5	14.7	17.1	13.5	12.7	12.8
Proficient	13.9	9.3	7.3	13.5	13.8	15.1	9.3	8.1	11.2
Advanced	17.5	15.5	14.2	13.9	11.8	14.7	24.2	11.5	16.3

SOURCE: D.H. McLaughlin, "Validity of the 1992 NAEP Achievement-Level-Setting Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 102.

process went on. This result (which may have been related to the feedback judges received about the desired grade-to-grade progression) had the effect of producing less overlap in the final standards between grades.

There was great variation in individual judges' recommended cutscores around the group average; and, with few exceptions, individual variation did not diminish much by Round 3. By Round 3, the standard deviations of judges' cutpoints ranged from 7.3 to 18.5 points on the NAEP scale. These numbers imply that even at the end of the process, the judges disagreed substantially about where the cutpoints should be set. A standard deviation of 14 points means that it takes a range of 28 points to represent even the middle two-thirds of the judges. On the NAEP scale, the within-grade standard deviation is approximately 40 points. Therefore, standards set by even the middle group of judges differed by as much as the actual performance of students at roughly the 25th and 50th percentiles. The change in standard deviations from Round 1 to Round 3 was generally small; in a few cases judges actually disagreed more at Round 3 than at the beginning. In some cases, the pattern was more like what would be expected if judges were changing their ratings on the basis of agreements with other group members. Observers of the process in reading noted, for example, that one of the fourth-grade groups discussed each item to come to agreement before filling out the forms; indeed, the data in table 3.9R show the expected reduction in variance for this grade. Overall, *there was little evidence that judges were reaching an authentic consensus or converging on an agreed-upon standard. Instead, these data suggest that the Angoff procedure merely averaged individual opinions, with minor adjustments to conform to item p-values and to improve grade-to-grade coherence.*

Analyses were also conducted to examine the effects on cutscores of having raters from different occupational and population groups, namely teachers, other educators, and representatives of the general public; white versus all other races; and men versus women. These data, in the form of intended cutscores reported separately for each group, are presented in tables 3.10M and 3.10R. There were no systematic differences in reading or mathematics associated with membership in any of these groups. Of course, the different grade-level groups, also shown in tables 3.10M and 3.10R, established increasingly higher cutscores for higher grade levels. Contrary to the expectation in some quarters that teachers might set lower standards than members of the public who are expert in the subject area, teachers consistently set slightly higher standards than other groups in reading. In mathematics, teacher standards were very similar to those set by other groups except at the basic level, where their expectations were lower, by 10 points, than the cutpoint set by members of the general public. *Given the large variation in judges' individual ratings and lack of group differences, it appears that the variation in judges' opinions was due to individual perspectives and not to systematic group effects.*

These numerical results on the effects of group membership are consistent with findings from the reading observational study, which can be summarized as follows. The roles taken by educators and noneducators and the amount of group interaction depended largely on individual personalities and the mix of individuals assigned to each table. Individuals held very diverse views about curriculum expectations in reading that were often at odds with the NAEP Reading Framework. There was no clear process to facilitate or develop consensus. Participants expressed a great deal of frustration at having to come together with such a diverse group and try to learn and respond to a new way of thinking about reading. "The most common criticism of the

Table 3.10M. ***Variations in mean mathematics cutpoint estimates between groups***

Group of Panelists	Basic	Proficient	Advanced
Fourth Grade	214.0	252.2	283.2
Eighth Grade	258.7	300.9	335.0
Twelfth Grade	292.8	336.3	365.8
Teachers	252.1	294.3	325.4
Other Educators	251.6	295.2	329.3
General Public	261.9	299.9	329.3
White	251.5	292.4	327.1
All Other Races	258.9	300.6	328.9
Women	255.2	297.7	328.8
Men	255.2	295.2	327.2

Table 3.10R. ***Variations in mean reading cutpoint estimates between groups***

Group of Panelists	Basic	Proficient	Advanced
Fourth Grade	208.2	241.3	273.4
Eighth Grade	238.7	279.3	324.9
Twelfth Grade	263.4	301.5	344.4
Teachers	241.6	277.9	317.3
Other Educators	235.1	273.7	310.6
General Public	233.6	270.5	314.8
White	235.2	275.0	316.7
All Other Races	238.4	273.0	311.8
Women	236.0	273.0	314.1
Men	237.6	275.0	314.4

SOURCE: D.H. McLaughlin, "Validity of the 1992 NAEP Achievement-Level-Setting Process," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 105.

level-setting process was that it moved too quickly to allow careful consideration of each step, to allow for understanding and reflection on the part of participants, to allow groups to coalesce, and to arrive at consensus." [13]

When interviewed about how they used the information provided during different rounds, reading participants said that they changed their item ratings in response to feedback. However, judges' interpretations of actual student performance and interjudge consistency information were apparently idiosyncratic. Therefore, common feedback did not dramatically improve agreement in judgments. Furthermore, there was very little discussion of the kind that would facilitate group consensus—with the exception of one fourth-grade group that discussed and reached consensus on every item. The authors of the reading process study concluded that "The St. Louis level-setting process cutpoints were arrived at by averaging across large numbers of ratings rather than through discussion of substantive matters and arrival at consensus."[14] *The Panel concludes similarly that the achievement levels in both reading and mathematics do not appear to reflect consensus standards or common expectations.*

Critique of the Angoff Method

The internal validity studies address questions about how well the modified Angoff procedure was implemented and whether judges' ratings were logically consistent. These studies also suggest more fundamental questions about whether the Angoff method or any other item-judgment method is adequate for conceptualizing subject-matter standards.

The Angoff method was invented initially in the context of minimum-competency proficiency judgments. Its use makes the most sense when there is a clear decision that must be made, such as licensing or certification of professionals (e.g., physicians or air traffic controllers), and the test items are designed to measure necessary knowledge or skills for that profession. Judges are asked to conceptualize individuals who are minimally competent to be certified and estimate the probability that such individuals would answer an item correctly. For example, a physician candidate must know when to administer insulin to a diabetic patient. Therefore, we might expect that the probability of a correct response to an item tapping this knowledge would be nearly 1.0 even for minimally competent physicians. The reasonableness of using the Angoff method becomes more and more murky when items span a range of content and difficulty, and judgments thereby lose their "must know to be minimally competent" quality.

The Angoff Method Requires an Unreasonable Cognitive Task

To cope with the problem that items on a continuous and heterogeneous scale cannot be classified as "must-pass," the modified Angoff method asks judges to imagine a group of examinees at the borderline of a group defined by an achievement-level

[13] Pearson and DeStefano, op. cit., 23.

[14] Ibid., 41.

description and to estimate what proportion of the borderline group would get the item right. This is an unreasonable cognitive task because judges have no basis for making such judgments. They nonetheless respond to the demand characteristics of the situation but in idiosyncratic ways, relying (at least in the case of reading) on personal experience, opinion, and intuition. It should be no surprise that judges cannot make these judgments very well. Unfortunately, we know of no previous studies that have used the kinds of analyses available here to investigate the internal consistency of judges' ratings. Judges are reasonably good at rank-ordering items by difficulty; in addition, the Angoff procedure provides them with data on item difficulty. *What judges cannot do conceptually is make sensible decisions about differences in probabilities within some plausible range.* For example, if judges are rating a difficult item, paying close attention to the definition of advanced performance, it is still a tossup whether they should say that 60 percent of borderline advanced or 80 percent of borderline advanced should get the item right. The large variation in individual judges' ratings reported in tables 3.9M and 3.9R is caused most probably by a combination of differences in interpreting the levels and descriptions (or applying their own definitions) and the difficulties in picking p-values. Averaging p-values across judges does not solve the conceptual problem; it only disguises it.

Limitation of the Angoff Procedure Not Merely Technical

If the limitations of the Angoff procedure are seen as technical problems leading to unreliable judgments, the tendency is to try to fix the problem by increasing the number of panelists to get more stable results or by tinkering with the details of the procedures. However, the problems with the Angoff methodology are not merely technical. *By focusing on items one at a time, for example, the method prevents judges from arriving at an integrated conceptualization of what performance at each of the levels should look like.* It never allows consideration of the combination of skills that should constitute proficient or advanced performance. Must an individual get all advanced items correct to be considered advanced? Should advanced global performance in mathematics require being advanced in all subdomains of mathematics, or could it mean being advanced in three areas and proficient in two? How can an advanced achievement level be established when important aspects of performance at that level are represented by only a few items or none at all?

Consensus Not Facilitated

Bringing together a large and diverse group of educators and noneducators to set standards has the appearance of a consensus process. In reality, however, Angoff panelists never get to discuss the cutpoints they are setting. Judges turn in individual p-values, and an average cutpoint is determined statistically. Judges do not have a chance to evaluate this end product or to deliberate about whether it is set too high or too low. They are given quantitative feedback and shown the rising pattern of cutpoints across grades, but they do not receive substantive feedback. They do not get to see, for example, the kinds of items that appear on either side of the cutpoint on the NAEP scale. Although not intended as a standard-setting method, the strategy used

by the expert validation teams, which is based on item mapping, enabled judges to participate in more substantive discussions about the nature of performance expected at each level. (See chapter 4.) Validation team members could also discuss directly whether they were satisfied with where they had drawn the line differentiating items of different types. The whole-booklet study provided an even better method for judges to apply verbal descriptions of the levels to more integrated examples of student work.

The expense and effort of the Angoff approach are misdirected because judgments of the kind required are beyond the capacity of panelists to make and because the procedures and time limits prevent panelists from engaging in debate over the most critical issues. The Angoff panelists were encouraged to participate in substantive discussions when they were working on the definitions but ironically, not when they were setting the standards.

Adjustments to the Final Cutpoints

In the Angoff procedure used by NAGB, recommended cutpoints on the NAEP scale were obtained for each achievement level by combining estimated p-values across all items and judges. The specific procedure for combining results was based on the NAEP scaling method, but can be thought of as a statistical averaging process. Thus, the recommended cutpoints were essentially the average of the individual judges' recommendations. These suggested cutpoints were then reported to NAGB for approval. In the case of mathematics, NAGB used a "standard error of measurement" based on variation in judges' estimates between two half-samples and lowered each cutpoint by one standard error. The rationale for this decision was "to give students the benefit of the doubt" by allowing students within one standard error of each cutpoint to be counted in the higher category. In reading, NAGB did not follow the same procedure. The cutpoints recommended by the St. Louis panelists were adopted without adjustment.

Given that the Angoff procedure does not produce a "true" standard, there is ample justification in the technical literature for revising proposed cutpoints in response to additional information. For example, NAGB might well have decided to set lower cutpoints in mathematics (at least at grade 12 advanced) given the number of students qualifying for college credit on Advanced Placement examinations. (These data are presented in chapter 4.) However, the particular argument used for "adjusting" the cutpoints in mathematics was unfortunate because it leaves NAGB and NAEP results vulnerable to criticism in two respects (although both complaints are unlikely to come from the same audience). First, the question arises, why not make the same "correction" in reading as in mathematics? Second, the use of a statistical correction creates a false impression of scientific accuracy. In the technical literature, standard-setting methods are acknowledged to be arbitrary and fallible. It is understood that average judgments are not estimates of a population parameter. However, making "a one-standard-error adjustment" treats the original cutpoint as if there were a parameter. This disguises the adjustment as a scientific decision and makes it appear that a true parameter is being estimated with fine tuning required to correct for only one source of error.

The final descriptions for the reading achievement levels were developed in three stages. The initial descriptions were developed by the consensus panelists in St. Louis and then revised at a subsequent meeting in San Diego. A third and final version was rewritten by NAGB staff or consultants. An example of how the reading descriptions changed from version to version is shown in figure 3.6 for the 12th-grade advanced level. As part of the reading process evaluation, Pearson and DeStefano convened a panel of reading experts to review the three versions of the descriptions. The experts concluded that the St. Louis and San Diego descriptions were quite similar. This conclusion is consistent with observational data. San Diego participants "felt constrained by warnings from ACT staff that the descriptors could not change substantively from those developed in St. Louis and by the protests of St. Louis participants who were heavily invested in the earlier versions."[15] The expert reading panel also concluded that the first two versions of the descriptions did not adequately reflect the Reading Framework. The third version was judged to be more congruent with the framework. However, this alignment was accomplished by significantly changing the achievement-level definitions that had been used in making item judgments and in setting the achievement-level cutpoints. In correcting one mismatch, a new mismatch was created which may invalidate the agreed-upon cutpoints.

The Panel does not have such extensive observational and participant survey data available in mathematics as for reading. However, some of the same difficulties with descriptions arose with the mathematics achievement levels. In mathematics, there were two versions of the level descriptions, one developed at the St. Louis level-setting meeting and one produced at Nantucket 4 months later. (The Nantucket version with minor editing became the final version.) A team of mathematics experts convened for one of the external validation studies reported in chapter 4 was also asked to review the two versions of the mathematics descriptions. They concluded that the two versions "differed in nontrivial ways, with the St. Louis version more closely matching the 1992 NAEP Mathematics Framework and items, and the final version better representing mathematics achievement aspirations that match contemporary thinking."[16]

ACT staff were aware that changing the level descriptions after the judgment process could invalidate the level cutpoints. They administered a questionnaire to ask St. Louis participants about how the changes in descriptions might have affected item ratings. However, this is an empirical question that cannot be answered adequately by opinion data. The Panel modified both of its experimental studies, conducted on eighth-grade mathematics, to attempt to examine how changing the descriptions might change the levels. The results of these experimental comparisons were inconclusive. In one study, slightly and nonsignificantly higher levels were created in response to the St. Louis descriptions. In the other study, one of the three levels was significantly different with higher levels being set with the final Nantucket descriptions. Because

[15] D. Pearson and L. DeStefano, "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report Two: An Analysis of the Achievement-Level Descriptions," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 193.

[16] E.A. Silver and P.A. Kenney, "Expert Panel Review of the 1992 NAEP Mathematics Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 240.

Figure 3.6. ***Examples of three versions of the 1992 NAEP reading achievement-level descriptions***

Basic Performance Grade 12–St. Louis Version

Basic performance in reading should include:

- Explaining the main idea of a text
- Describing the main purpose in reading a selection
- Recognizing the significance of details from a reading in order to support a conclusion or perform a task
- Applying information gathered from reading to meet an objective or support a conclusion
- Explaining the basic elements of an author's literary devices

Basic Performance Grade 12–San Diego Version

Basic performance in reading should include:

- Explaining the main idea, theme, or purpose of a text
- Describing the main purpose for reading a selection
- Recognizing the significance of details from a reading in order to support a conclusion or perform task
- Applying the information gathered from reading to meet an objective or support a conclusion
- Identifying and explaining the basic elements of an author's literary devices
- Making logical connections between a text and personal knowledge and experience
- Maintaining a focus over the entirety of a story/informational text

Basic Performance Grade 12–Final Version

Twelfth grade students performing at the basic level *should be able to demonstrate an overall understanding and make some interpretations of the text.* When reading text appropriate to 12th grade, they *should be able to identify and relate aspects of the text to its overall meaning, recognize interpretations, make connections among and relate ideas in the text to their personal experiences, and draw conclusions.* They *should be able to identify elements of an author's style.*

For example, when reading literary text, 12th-grade students should be able to explain the theme, support their conclusions with information from the text, and make connections between aspects of the text and their own experiences.

When reading informational text, basic-level 12th graders should be able to explain the main idea or purpose of a selection and use text information to support a conclusion or make a point. They should be able to make logical connections between the ideas in the text and their own background knowledge.

When reading practical text, they should be able to explain its purpose and the significance of specific details or steps.

SOURCE: D. Pearson and L. DeStefano, "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report Two: An Analysis of the Achievement-Level Descriptions," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 210.

the findings from these small sample studies were in different directions and applied to only one grade level in one subject area, the Panel could not draw any general conclusion about the magnitude of changes likely to occur in cutpoints when substantive descriptions are changed. However, since one cannot know how different the cutpoints might have been if judges had been using the final descriptions to make their item judgments, *it is logically indefensible to report using cutpoints that are potentially inconsistent with the narrative descriptions.* In the future, a process for developing performance standards should allow for revision and iteration so that quantitative cutpoints, verbal descriptions, and exemplar tasks are developed in a coherent and integrated manner. This issue is discussed again in chapter 5.

In addition to the need for congruence between level descriptions and level cutpoints, the foregoing experiences suggest other issues that must be addressed in the development of national performance standards. What is the role of experts? How should pronounced differences of opinion be resolved or represented? Although several factors contributed to the systematic differences between the first and second versions of the descriptions in both reading and mathematics, one explanation is that research-oriented experts who traditionally participate along with other educators in development of NAEP frameworks were not represented at the initial St. Louis meetings. They were, however, invited to the followup meetings in San Diego and Nantucket, respectively. As a group, educators who had not participated in St. Louis tended to express dissatisfaction with the earlier versions of the descriptions because they did not reflect contemporary conceptions of subject matter. In the case of mathematics, the newcomers exerted significant pressure to change the descriptions. In reading, a third version created after San Diego was more in keeping with the conceptions advocated by reading experts.

More deliberation is needed about the appropriate role for experts, about the intended relation between the descriptions and frameworks, and about better means to resolve or respect divergent views about subject matter. For example, with more time and with the full range of experts participating from the start, there might be more opportunity for review of the frameworks and for discussion of the rationale behind current content standards. In chapter 5, the Panel also discusses the possibility that differences between traditional and more contemporary content goals may sometimes be so great that they cannot be merged but should be assessed and reported separately.

Summary

The process of developing achievement levels involved two distinct tasks: creating subject- and grade-specific descriptions for each level and identifying cutscores. To evaluate the process, the Panel commissioned a series of studies that included reanalyses of the item-rating data collected during the 1992 standard-setting meetings, observations of the process in reading combined with interviews and written surveys administered to participants, reviews of different versions of the descriptions by content experts, and independent experimental studies of the judgment process. Key findings are summarized as follows:

1. ***In reading, the initial achievement-level descriptions and item judgments were inappropriately influenced more by panelists' personal experiences and opinions than by the Reading Framework.***

The reading evaluation study identified several serious problems with the development of the descriptions and subsequently in the way that descriptions were used to rate items. (1) Most participants were unfamiliar with the Reading Framework and relied instead on personal definitions of reading and personal experience to describe achievement levels. (2) The initial descriptions were influenced more by assessment items than the framework. This contributed to the misalignment of the descriptions and the framework because the current exercise pool does not adequately reflect all aspects of the Reading Framework. (3) Many participants reported using referent groups from their own experience to make item judgments rather than using the descriptions.

2. The process for developing the descriptions in reading and mathematics was inadequate because it did not ensure that final descriptions were agreed upon before attempting to set cutscores.

For a variety of reasons, the achievement-level descriptions in both reading and mathematics were revised significantly after the initial level-setting meeting. In the case of reading, the final descriptions were more closely aligned with the conception of reading in the framework. In mathematics, the final levels moved away from the mathematics framework toward a definition of mathematical abilities more consistent with current thinking in the field. In both cases, the changes were substantial enough to raise serious validity questions. The final descriptions may not be valid for describing the assessment and may not correspond to the cutpoints determined on the basis of earlier definitions. These issues, which bear on the validity of interpretations made from score reports, are discussed further in chapter 4.

3. The 1992 cutpoints set in reading and mathematics are indefensible because of large internal inconsistencies.

By reanalyzing data from the judgment process itself, it was possible to evaluate the internal consistency of individual judges' ratings and to determine which item features affected consistency. We had expected, for example, that judges might reasonably set higher standards (relative to empirical item difficulties) for aspects of content that are not currently being taught. Surprisingly, judges did not vary their ratings in response to substantive dimensions of the assessment. The Panel did, however, find large internal inconsistencies in judges' ratings in response to item features, such as right-wrong versus extended-response items, multiple-choice versus short-answer items, and easy versus hard items. These findings suggest that judges are unable to envision hypothetical borderline examinees and estimate p-values as required by the Angoff procedure.

4. The Panel concludes that the Angoff procedure is fundamentally flawed because it depends on cognitive judgments that are virtually impossible to make.

Although the Angoff procedure has been identified in the research literature as a popular standard-setting method, it is fair to say that data of the kind available to the Panel, especially estimates of judges' internal consistency, have not been examined previously. Findings from this particular application may not generalize to other testing situations. However, these findings are so dramatic and consistent that the adequacy of the Angoff method in other contexts should not be presumed until studies of this type are undertaken.

Based on logical analysis and findings from an experiment, where cutscores set using item judgments were compared to cutscores set using complete booklets of students' responses, *the Panel concludes further that item-by-item methods are inadequate for allowing judges to develop integrated conceptions of performance standards.*

5. The process used by NAGB did not facilitate the development of consensus, either in developing descriptions or in setting cutscores.

Observations of the reading process and interview responses suggest that participants were frustrated by divergent views and did not have sufficient time for appropriate discussions and interactions to develop shared understandings of subject-matter expectations. The statistical data suggest that feedback to participants improved the internal consistency of individuals' ratings and the coherence of the levels across grade levels. However, there was little evidence that the final levels represented a consensus among participants. The cutpoints changed very little from Round 1 to Round 3, and the variation among judges' personal recommended cutscores remained very large, even at Round 3. Thus, it cannot be claimed that the current levels are the product of a consensus process.

6. The decision to adjust cutscores in mathematics by "one standard error" was misleading.

Although other evidence should be brought to bear to adjust the initial cutscores, the Panel criticizes the decision to adjust the cutscores in mathematics "by one standard error" because doing so implies statistical precision in the recommended cutscores that is unwarranted.

The 1992 achievement levels for reading and mathematics were based on flawed procedures and therefore cannot be defended as the basis for reporting NAEP results. In chapter 5, the Panel considers procedures for developing performance standards that would redress deficiencies in the current process.

4 Validity and Reasonableness of the Achievement Levels in Reading and Mathematics

The Panel's evaluation studies were guided by an overarching research question. Will the use of the 1992 NAEP achievement levels lead to valid interpretations of NAEP results? In chapter 3, the Panel discussed the processes for setting the level cutpoints and for developing the level descriptions. Those evaluation studies are referred to collectively as the internal process studies. This chapter presents a series of studies focused on the outcomes of that process. Two major outcomes are considered: the achievement-level cutscores and the achievement-level descriptions. The first major section of this chapter, External Validity Comparisons, presents studies bearing on the reasonableness of the cutscores themselves. The second major section, Adequacy of Level Descriptions and Exemplar Items, addresses the validity of the accompanying descriptions and exemplar items, which give meaning to the terms *basic, proficient,* and *advanced.*

The studies reported in this chapter were designed and conducted independently of those in chapter 3, but they cannot be interpreted independently of one another. Internal and external evaluations of the achievement levels are analogous to the conjoint requirements for test reliability and validity. Reliability is a necessary but not sufficient condition for test validity. Analogously, internally consistent and sound processes are prerequisites for defensible standards, although external validity evidence is still required.

Having found that the process for setting the achievement levels was fundamentally flawed, it may seem odd that the Panel now goes on to evaluate the external validity of the outcomes of those processes. An explanation is in order. First, the external evaluation studies were planned with the expectation that the 1992 internal process would prove reasonable and supportable, or that problems would be relatively minor. Because resources were committed to these investigations, it is important that the data be made available to as wide an audience as possible. Second, the studies provide valuable insights about how the reasonableness of achievement levels might be judged even if in this case they cannot be used to "confirm" the reasonableness and external validity of the 1992 levels.

The first section of this chapter describes results of the external comparison studies that were designed to help judge the reasonableness of the achievement-level cutscores:

- Teacher classifications of students according to the NAEP achievement-level descriptions, based on their classroom work;
- Researcher classifications of a subset of the same students based on individually administered assessments, again according to the NAEP achievement-level descriptions;

- Comparisons to results of the Scholastic Aptitude Test (SAT) and Advanced Placement (AP) examinations;
- International comparisons in mathematics;
- Comparisons to results of the Kentucky state assessment;
- Examination of the pattern of cutpoints and student performance across grades; and
- Matching of achievement-level descriptions to the NAEP performance scales by content experts.

None of these external comparisons was expected to yield an absolute validity criterion. Rather, each of these studies provides external evidence of the numbers of students who meet the described standards for advanced, proficient, and basic performance and therefore helps to evaluate whether the achievement-level cutscores are making reasonable classifications. For example, the teacher-based and researcher-based studies are intended to identify students meeting each of the achievement levels using more extensive evidence of their performance than is possible from the NAEP assessment alone. Do teacher and researcher classifications of students corroborate the achievement-level cutscores? Similarly, AP examinations are well known as measures of advanced achievement. Students who score well on AP tests meet the verbal definition of advanced; does the proportion of students also align with the achievement-level cutpoint for 12th-grade advanced?

The second part of the chapter addresses the content validity of the achievement-level descriptions and of exemplar items. In mathematics, the Panel presents judgments from content experts already highly familiar with the NAEP Mathematics Framework, the exercise pool, and different versions of the achievement-level descriptions, and also summarizes results from an independent content analysis (by a second group of experts) of the grade 4 mathematics pool relative to the NCTM Standards.

In reading, the chapter covers evaluations of the level descriptions and exemplar items by an expert panel steeped in the NAEP reading assessment, comparable to the parallel group in mathematics, and by an independent group of reading researchers. For both reading and mathematics, an analysis of the "should versus can" interpretations of achievement-level results is also presented. All of these studies were designed to probe more deeply the alignments among item pools, NAEP frameworks, achievement-level descriptions, illustrative items, and discipline-based conceptions of content standards.

External Validity Comparisons—Are the Cutpoints Set at Reasonable Levels?

This section reports the findings of seven studies or data comparisons bearing on the question of whether the 1992 achievement levels set in reading and mathematics are appropriate, too high, or too low. Some of these studies are much more rigorous or

extensive than others. Therefore, the Panel has not simply counted the number of studies that say the standards are "too high" and averaged their results with studies that suggest that the standards are "too low." Rather, the Panel weighed the evidence, giving greater emphasis to findings that are consistent across grades and subject areas and to findings for which there are logical explanations. For example, the Panel gave the greatest attention to teacher judgments when they were confirmed by judgments from independent researchers and when they led to consistent results across grades for both reading and mathematics. The Panel gave less attention to the cutscore results from the content-expert, item-mapping studies because of discoveries made after-the-fact about the effect of the item-mapping statistical criterion employed. However, important insights about the item-judgment process were gained from these latter studies, vis-a-vis changes in descriptions and the effect of having an item pool with too few advanced items.

Evidence of Student Proficiency from Contrasting-Groups Studies

A classic method for gathering validity data in the standard-setting literature is the "contrasting-groups" study design. In fact, NAGB staff had originally recommended that judgmental methods be used to establish standards and that empirical methods such as contrasting-groups studies be used as "a verification procedure."[1] The contrasting-groups approach is both a source of validity evidence and a standard-setting method. Like other standard-setting procedures, it does not produce a "true" standard, but it does provide important information about the reasonableness of cutpoints given nontest evidence of student performance.

To implement the contrasting-groups design, judges or raters must have knowledge of what students can do in the domain measured by the test. In the case of the NAEP achievement levels, judges must be able to apply the definitions of advanced, proficient, and basic performance and to classify students into the appropriate categories. Then, in order to evaluate the correspondence between external judgments of students' proficiency and NAEP classifications based on the achievement levels, the judges' classifications must be translated into implied cutscores on the NAEP scale. The next steps, therefore, are to obtain NAEP scores for the same students, and to apply a statistical procedure that identifies cutscores which come the closest to classifying students the same way on NAEP as they are classified by the judges.

Design of the Panel's contrasting-groups studies. Contrasting-groups studies were conducted in reading and mathematics at both grades 4 and 8.[2] These studies required the collection of new data on reasonably large samples of students who had also taken the NAEP assessments. The data were collected in conjunction with the 1993 field trials for the 1994 assessments in reading and mathematics; and the studies, which involved 5,035 students in 171 schools, would not have been possible without extensive cooperation and support from the NAEP contractors. To connect results on

[1] Roy Truby, "Staff Paper on Setting Goals for the National Assessment" (Washington, D.C.: National Assessment Governing Board, December 8, 1989), 19.

[2] D.H. McLaughlin et al., "Comparison of Teachers' and Researchers' Ratings of Students' Performance in Mathematics and Reading with NAEP Measurement of Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993).

new field-test items to the NAEP 1992 achievement levels, 1992 NAEP booklets were administered randomly within the participating classrooms. The four samples of students in reading and mathematics and for grades 4 and 8 ranged in size from 1,078 to 1,415.

To identify the achievement level of each student, a student's teacher (for mathematics or reading) was provided with the appropriate NAEP achievement-level descriptions and was asked to classify the student's classroom performance as advanced, proficient, basic, or below basic. Thorough instructions were included, along with the detailed achievement-level descriptions. Teachers coded their classification (A, P, B, BB) on an insert that was added to the standard NAEP "teacher questionnaire" given to teachers of participating students. In appreciation of the extra time spent by teachers on this validation task, each participating school received $100 to spend on instructional supplies. Indications such as completeness of data and telephone calls by teachers to an 800-number to ask about the details of the definitions suggest that teachers completed the task conscientiously.

Researcher validation of teacher ratings. Teachers have extensive knowledge of what students in their classrooms are able to do. A significant concern in using teacher ratings, however, is that each teacher might interpret the definitions differently or that teachers might resort to "normative" ratings (i.e., rating their best students as advanced regardless of whether they can do what is described in the achievement levels). In the instructions, teachers were urged to pay attention to the definitions and not to rate children in relation to others in the class. They were also told explicitly that it would probably not be appropriate to classify children as advanced if topics specified in the definitions had not yet been covered in the curriculum. There is strong evidence that teachers accepted these instructions; specifically, in 63 of the 171 schools, teachers identified no participating students as advanced.

To check on the accuracy of teachers' ratings, followup individual assessments were administered to students in a random subsample of schools (approximately 15 of 45 schools in each of the four studies). Interview protocols were developed to allow researchers to work one-on-one with students to elicit evidence of students' level of performance in reading or mathematics. In mathematics, for example, problems that would address the five NAEP content strands were selected or adapted from the NCTM Standards document. A sample mathematics problem, one of six used to assess eighth-grade students, is shown in figure 4.1. Each "problem" was, in fact, multiple problems. If students could do the first part of a problem, they were asked to go on and extend their answer and to do additional parts of the problem. When students could not respond, they were given prompts that would make the problem easier. Commensurate adjustments were then made in how the problem was scored. For the problem in figure 4.1, a student would have to complete all parts without any help to be considered advanced on that problem. Scoring rules were also developed for combining scores across problems. For example, in mathematics students had to be advanced on all problems, or proficient on one and advanced on the others, to be classified as advanced overall. Thus, researcher classifications were intended to be stringent and consistent with the verbal descriptions.

Results in reading. Results from the contrasting-groups studies are shown in tables 4.1 and 4.2 for reading and in tables 4.3 and 4.4 for mathematics. For each study, cutpoints and percentages at or above each level are shown for (1) the official

Figure 4.1. ***One sample problem from the eighth-grade mathematics protocol used by researchers in individual assessments of student proficiency***

Problem 1-1. Here is a table. I am going to ask you some questions about it.

Waffles To Go				
waffles	cups mix	cups milk	eggs	tbs. oil
4	2	1.5	1	2
6	3	2.25	1.5	3

How many waffles can you make with 2 cups of mix? (Must be able to read the table to proceed.)

1-2. To make 12 waffles, how much of each ingredient is needed?

1-3. (Give the student blank graph.) Sketch a graph of the function which shows how many waffles you can make from each number of cups of mix.

1-4. Write an equation that shows the relation between the number of waffles and the number of cups of mix.

Note: If students had difficulty on any part of a problem, they were provided with prompts designed to make the problem easier, with commensurate reductions made in the scoring of the student's level. To be considered advanced on this problem, an eighth-grade student would have to have completed all four parts of the question as given without additional prompts.

* Questions were based on a spreadsheet example given in National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (Reston, VA: Author, 1989).

achievement levels, (2) teacher classifications, and (3) the researcher classifications.[3] *In reading at both grades 4 and 8, teachers and researchers applying the NAEP definitions identified many more students as advanced, proficient, and basic than were identified using the NAEP achievement-level cutpoints.* For example, at grade 4, the achievement-level cutpoints would classify only 0.7 percent of students as advanced. Teachers identified 11 percent of the same sample of students as advanced, and researchers said that 15 percent were advanced. NAEP achievement levels classified only 11 percent as performing proficiently or above. In contrast, teachers found a total of 39 percent to be proficient or above in reading, and researchers judged that 30 percent demonstrated proficiency in a one-on-one assessment.

Results in mathematics. Results in mathematics followed a similar pattern, especially for fourth grade. At grade 4, both teachers and researchers identified more students at each of the achievement levels than were identified by the NAEP achievement levels. Two percent were classified as advanced using the achievement levels, whereas teachers identified 8 percent of students as advanced in mathematics and researchers confirmed that 6 percent were advanced. The achievement levels classified 12 percent of students as proficient compared to teachers and researchers who classified 31 and 38 percent respectively as proficient. Grade 8 mathematics was the only one of the four studies where this pattern was not perfectly consistent. For eighth-grade mathematics, teachers again identified more students at each of the three levels than were so classified by the achievement levels. However, teachers' classifications were not always confirmed by researchers. Instead, researchers' results agreed more with the achievement-level results at the basic and advanced levels, and were intermediate between the teacher and achievement-level results for the proficient level.

Correspondence between teacher and researcher ratings. It is important to emphasize that the higher ratings of students' performance by teachers compared to the achievement-level classifications do not appear to be the result of leniency or lower standards on the part of teachers given that independent ratings by researchers led to similar results in 10 of 12 comparisons. In separate analyses of the correspondence between teacher and researcher ratings, there was no evidence that teachers gave systematically higher or lower ratings than researchers in three of the four studies. Only for eighth-grade mathematics were teachers' cutpoints systematically lower than researchers' cutpoints. The cross-tabulations of teachers' judgments about classroom performance and researchers' judgments about performance in a one-on-one assessment are shown in tables 4.5 and 4.6 for reading and 4.7 and 4.8 for mathematics. In reading, the percent agreement between ratings by teachers and researchers, adjusted for different marginal distributions, was 75 percent and 66 percent for grades 4 and 8, respectively. In mathematics, the percent agreements were somewhat smaller, 44 percent and 59 percent.

Summary of results from contrasting-groups studies. In reading, the results for both grades 4 and 8 follow a consistent pattern which suggests that the achievement-level cutpoints have been set systematically too high. The sizes of these effects are substantial, with the current cutpoints leading to serious underreporting of students who are functioning at the advanced and proficient levels in reading. These findings,

[3] The percentages at or above the official cutpoints are not the same as for the national sample because these data are from a sample drawn from the 1993 *field-test*. At grade 8 these field-test data are quite similar to the national sample. However, at grade 4, particularly in reading, the field-test sample appears to be less able than the national sample, with lower percentages of students scoring at each of the levels. Differences between the national and study samples do not threaten the validity of the study so long as there are sufficient numbers of students in each category. *The relevant comparisons are among the different cutpoints and percentages for the study samples.*

Table 4.1. ***Fourth-grade reading achievement-level cutpoints and percentages of field-test students achieving, based on three sources***

	Criterion Based on NAEP, as set by the Governing Board		Criterion Based on Teachers' Ratings of Classroom Performance		Criterion Based on Researchers' Ratings of Performance	
	Cutpoint	Percent	Cutpoint	Percent	Cutpoint	Percent
Basic	212	42 ± 3	171 ± 3	80 ± 3	188 ± 4	68 ± 5
Proficient	243	11 ± 1	214 ± 2	39 ± 3	226 ± 4	30 ± 4
Advanced	275	0.7 ± .1	249 ± 3	11 ± 2	245 ± 4	15 ± 4

Table 4.2. ***Eighth-grade reading achievement-level cutpoints and percentages of field-test students achieving, based on three sources***

	Criterion Based on NAEP, as set by the Governing Board		Criterion Based on Teachers' Ratings of Classroom Performance		Criterion Based on Researchers' Ratings of Performance	
	Cutpoint	Percent	Cutpoint	Percent	Cutpoint	Percent
Basic	244	78 ± 2	223 ± 4	86 ± 2	236 ± 6	83 ± 4
Proficient	283	31 ± 2	273 ± 3	46 ± 4	273 ± 4	49 ± 5
Advanced	328	1.5 ± .4	303 ± 3	15 ± 2	307 ± 5	13 ± 3

Note: Cutpoints based on ratings minimize the asymmetry in misclassifications between categories. Standard error estimates are based on observed between-school variation. The percentages for the researchers' ratings are based on the subsample of students rated by the researchers.

SOURCE: D.H. McLaughlin et al., "Comparison of Teachers' and Researchers' Ratings of Students' Performance in Mathematics and Reading with NAEP Measurement of Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 293.

Table 4.3. ***Fourth-grade math achievement-level cutpoints and percentages of field-test students achieving, based on three sources***

	Criterion Based on NAEP, as set by the Governing Board		Criterion Based on Teachers' Ratings of Classroom Performance		Criterion Based on Researchers' Ratings of Performance	
	Cutpoint	Percent	Cutpoint	Percent	Cutpoint	Percent
Basic	211	40 ± 2	186 ± 2	84 ± 2	195 ± 4	73 ± 6
Proficient	248	12 ± 2	226 ± 4	31 ± 3	214 ± 4	38 ± 6
Advanced	280	2 ± .7	253 ± 4	8 ± 2	244 ± 7	6 ± 2

Table 4.4. ***Eighth-grade math achievement-level cutpoints and percentages of field-test students achieving, based on three sources***

	Criterion Based on NAEP, as set by the Governing Board		Criterion Based on Teachers' Ratings of Classroom Performance		Criterion Based on Researchers' Ratings of Performance	
	Cutpoint	Percent	Cutpoint	Percent	Cutpoint	Percent
Basic	256	70 ± 2	226 ± 5	89 ± 2	251 ± 6	83 ± 5
Proficient	294	27 ± 3	276 ± 5	50 ± 4	286 ± 4	40 ± 6
Advanced	331	5 ± 1	308 ± 4	15 ± 2	331 ± 6	5 ± 2

Note: Cutpoints based on ratings minimize the asymmetry in misclassifications between categories. Standard error estimates are based on observed between-school variation. The percentages for the researchers' ratings are based on the subsample of students rated by the researchers.

SOURCE: D.H. McLaughlin et al., "Comparison of Teachers' and Researchers' Ratings of Students' Performance in Mathematics and Reading with NAEP Measurement of Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 294.

Table 4.5. ***Frequencies of fourth-grade-student reading-performance classification by teachers and researchers***

		Classification of Performance by Researchers				
		Below Basic	Basic	Proficient	Advanced	Total
Classification of Performance by Teachers	Below Basic	17	6			23
	Basic	18	30	2	2	52
	Proficient	4	11	12	5	32
	Advanced			5	13	18
	Total	39	47	19	20	125

Table 4.6. ***Frequencies of eighth-grade-student reading-performance classification by teachers and researchers***

		Classification of Performance by Researchers				
		Below Basic	Basic	Proficient	Advanced	Total
Classification of Performance by Teachers	Below Basic	11	5	1		17
	Basic	7	22	6	3	38
	Proficient	1	5	21	3	30
	Advanced		1	9	8	18
	Total	19	33	37	14	103

SOURCE: D.H. McLaughlin et al., "Comparison of Teachers' and Researchers' Ratings of Students' Performance in Mathematics and Reading with NAEP Measurement of Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 312.

Table 4.7. ***Frequencies of fourth-grade-student mathematics-performance classification by teachers and researchers***

		Classification of Performance by Researchers				
		Below Basic	Basic	Proficient	Advanced	Total
Classification of Performance by Teachers	Below Basic	14	14	3		31
	Basic	13	16	12	2	43
	Proficient	9	12	12	5	38
	Advanced	5		5	3	13
	Total	41	42	32	10	125

Table 4.8. ***Frequencies of eighth-grade-student mathematics-performance classification by teachers and researchers***

		Classification of Performance by Researchers				
		Below Basic	Basic	Proficient	Advanced	Total
Classification of Performance by Teachers	Below Basic	6	7	2		15
	Basic	11	16	10	1	38
	Proficient	3	19	18	2	42
	Advanced	3	8	10	4	25
	Total	23	50	40	7	120

SOURCE: D.H. McLaughlin et al., "Comparison of Teachers' and Researchers' Ratings of Students' Performance in Mathematics and Reading with NAEP Measurement of Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 313.

based on more extended evidence of student work, are logically consistent with results in chapter 3 showing that item-by-item judgments led to higher cutpoints for advanced and proficient levels (hence finding fewer students in these categories) than when judges applied the same descriptions to whole booklets of student work.

In mathematics, the results are not so simple. Clearly, in comparison to teachers' judgments of classroom performance in both grades, the NAEP achievement-level cutpoints were too high; however, researchers would have departed from the achievement-level cutpoints for only one of the two grades.

In addition to the quantitative findings, the contrasting-groups studies are also useful as demonstrations of an alternative standard-setting methodology, or more properly, a methodology to collect standard-setting evidence as one part of a larger process. This alternative methodology appears to be superior to the modified Angoff procedure. Recall that in chapter 2, a distinction was presented between judgmental, or "test-centered" standard-setting methods (like the Angoff method) and empirical, or "examinee-centered" methods (like the contrasting-groups method). The latter are based on judgments about *actual* individual students rather than hypothetical populations of borderline examinees. Although based on real students, this method does not imply that the standards become "normative," because judges must identify students who meet the verbal descriptions of criterion performance. The teacher and researcher classification studies reported here demonstrate the feasibility of these methods for standard setting on NAEP, including the feasibility of collecting relevant data as part of field testing or full-scale assessments.

SAT and AP Examination Comparisons for 12th Graders

The SAT and AP test results are not available for a representative sample of U.S. 12th graders. Nonetheless both experts and policymakers have developed a great deal of familiarity with these tests both in terms of the content they measure and in terms of the performances (in high school and in college) of students who score at certain levels. For example, users of SAT scores have a sense of what a score of 600 on the Mathematics test "means." Therefore, it is informative to consider how results on the SAT and AP tests compare with the percentages of students judged to be advanced using the NAEP achievement levels.

Comparison of achievement levels with SAT results. Data are provided in table 4.9 showing the percentage of students scoring above 550 on the SAT Verbal and above 600 on the SAT Mathematics tests.[4] The percentages in the table are calculated by counting the number of SAT test-takers above the designated score and dividing by the number of high school graduates in that year. These percentages are *not* the percentage of test-takers above 550 and 600. Because test-takers are a select population, a much higher percentage of test-takers are above these levels. In 1992, for example, 14 percent of test-takers scored above 550 on the Verbal test, and 18 percent of test-takers scored above 600 on the Mathematics test. However, in order to make the

[4] The SAT Verbal test is known to be much more difficult than the Mathematics test. The College Board has, in fact, recently made the decision to rescale the Verbal test to make the score scales for the two measures more equivalent. For this reason, the Panel adopted the cutpoint of 550 on the Verbal test as roughly equivalent to 600 on the Mathematics test.

percentages comparable to the advanced percentages obtained on the NAEP achievement levels, it was necessary to estimate the percentage of all 12th graders scoring at these levels, not the percentage of test-takers. These estimates are crude and almost certainly underestimate the percentage of 12th-grade students capable of scoring at a given level because many college-bound high school students take the ACT and not the SAT.

Table 4.9. ***Percentage of high school graduates scoring above SAT score of 550 (Verbal) and 600 (Math) for 1990, 1991, 1992***

School Year Ending	Number of High School Graduates	SAT Scores			
		Verbal		Mathematics	
		Mean	Percent 550 or higher	Mean	Percent 600 or higher
1990	2,589,705[a]	424[a]	5.7[e]	476[a]	7.1
1991	2,506,517[a]	422[a]	5.8[d]	474[a]	7.0
1992	2,485,000[b]	423[c]	5.8[c]	476[c]	7.5

[a]SOURCE: *The Condition of Education*, NCES, 1992
[b]SOURCE: U.S. Department of Commerce, Bureau of Census
[c]SOURCE: *College Bound Seniors: 1992 Profile of SAT and Achievement Test Takers*
[d]SOURCE: *College Bound Seniors: 1991 Profile of SAT and Achievement Test Takers*
[e]SOURCE: *College Bound Seniors: 1990 Profile of SAT and Achievement Test Takers*

Scores of 550 and 600 are arbitrary cutpoints, but we have reason to believe that they represent reasonably high levels of achievement. For example, data provided by the College Board show that students who are in the top 10 percent of their high school class earn average SAT Verbal and Mathematics scores of 512 and 585, respectively. Students who expect to qualify for advanced placement in mathematics have an average SAT Mathematics score of 580. Students reporting advanced placement in English have an average SAT Verbal score of 500.

Actual items from past SAT exams in mathematics are shown in figure 4.2. As these items illustrate, content on the SAT is as difficult as the most advanced content presently included in NAEP for 12th grade. (For comparison, see the two advanced exemplar items used to report NAEP mathematics results for 1992, which are shown later in figure 4.3.) The Panel does not cite the SAT as an ideal measure of mathematics achievement; it falls short in representing the NCTM Standards necessarily because of the number of items administered in a short period of time, the lack of opportunity for students to explain their answers, and so forth. Nonetheless, SAT items do assess the NAEP content strands and do require mathematical reasoning. Students who do well on the kinds of problems illustrated in figure 4.2 are meeting the description for advanced 12th-grade performance in mathematics as well as they could meet it on NAEP. Therefore, the percentage of students earning high scores on the SAT tell us something important about the proportion of 12th graders who are advanced.

Figure 4.2. ***Typical items from the SAT mathematics test which are equal to or greater in difficulty than the 12th-grade advanced exemplar items***

$$\begin{array}{r} 1X3 \\ Y3 \\ +\ \ Z6 \\ \hline 312 \end{array}$$

1. In the addition problem above, which of the following could be the digit Z?

   I. 1
   II. 5
   III. 8

   (A) II only
   (B) III only
   (C) I and II only
   (D) II and III only
   (E) I, II, and III

2. What is the number of different pairs of parallel edges on a cubical wooden block?

   (A) 18 (B) 12 (C) 8 (D) 6 (E) 4

3. The average (arithmetic mean) of n numbers is 10 and the average of 5 of these numbers is 8. In terms of n, what is the average of the remaining numbers?

   (A) $\frac{10n + 40}{5}$

   (B) $\frac{10n - 40}{5}$

   (C) $\frac{40 - 10n}{5}$

   (D) $\frac{40 - 10n}{n - 5}$

   (E) $\frac{10n - 40}{n - 5}$

4. If x and y are two different integers and the product $35xy$ is the square of an integer, which of the following could be equal to xy?

   (A) 5
   (B) 70
   (C) 105
   (D) 140
   (E) 350

5. What is the perimeter of a square that has the same area as a circle that has circumference 1?

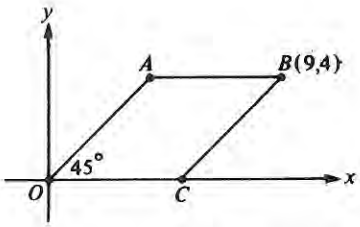
   (A) $4\sqrt{\pi}$

   (B) $\frac{2}{\sqrt{\pi}}$

   (C) $\frac{1}{2\sqrt{\pi}}$

   (D) $\frac{1}{\pi}$

   (E) 1



6. In the figure above, what is the perimeter of parallelogram $OABC$?

   (A) $10 + 4\sqrt{2}$
   (B) $10 + 8\sqrt{2}$
   (C) 18
   (D) 36
   (E) It cannot be determined from the information given.

7. If r is a positive integer and the ratio $\frac{r}{s}$ is $\frac{1}{3}$, which of the following could be the value of $\frac{r^2}{s}$?

   I. $\frac{1}{3}$

   II. 1

   III. $\frac{4}{3}$

   (A) None
   (B) I only
   (C) I and II only
   (D) I and III only
   (E) I, II, and III

SOURCE: The College Entrance Examination Board, *10 SATs*. (4th ed.) (New York: Author, 1990).

The SAT data suggest that the achievement-level cutpoints for 12th-grade advanced may have been set very high. In 1992, only 2 percent of U.S. 12th graders scored at the advanced level in mathematics using the NAEP achievement levels, whereas at least 7.5 percent of high school graduates scored at 600 or better on the SAT Mathematics test. (An even larger percentage would have been obtained if all students had taken the SAT.) Similarly, on the SAT Verbal test which measures vocabulary, verbal reasoning, and reading comprehension, 5.8 percent scored at a high level (550) whereas the NAEP achievement levels classified only 3.2 percent of 12th graders as advanced in reading. Again, the discrepancy between the two measures is most likely greater than it appears, given that the SAT figures are underestimates.

Comparison of achievement levels with AP results. Yet another imperfect but informative comparison considers the performance of high school students taking AP examinations. The AP program, under the auspices of the College Board, provides curricular materials to enable high schools to offer college-level course work to high school students. Examinations are then offered to evaluate students' level of achievement. These AP tests are often cited by advocates for education standards and assessments as positive examples of challenging, syllabus-driven examinations.

AP results for 1992 in calculus and English are reported in table 4.10.[5] The AP examinations are scored on a 1-5 scale. The number of students at each score point has been converted to a percentage of high school graduates using an estimate of the number of 1992 graduates provided by the U.S. Census. Most colleges award college credit for a score of 3 or better. By summing these three categories, we find that 2.5 percent of high school graduates score high enough to earn college credit in calculus and 3.8 percent qualify for college credit in English. These percentages are most surely underestimates given that AP examinations in any subject area are available in only 46 percent of U.S. secondary schools. In addition, scoring 3 or better on an AP exam might be considered to be "beyond advanced" because it signifies completion of college-level work rather than being prepared for college work.

Taken together, the approximate comparisons provided by the SAT and AP tests suggest that the NAEP advanced cutpoints in reading and mathematics have been set unreasonably high. From the perspective of these other tests, a higher percentage of 12th graders would be identified as advanced performers when compared to the percentages classified as advanced using the NAEP achievement levels.

International Comparisons in Eighth-Grade Mathematics

International comparisons have contributed to the sense that American students learn less than students in other nations and that expectations are set too low in this country. For example, in 1991, U.S. 13-year-olds ranked 14th of 15 in mathematics

[5] E. Hartka, "Comparisons of Student Performance on NAEP and Other Standardized Tests," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 370.

Table 4.10. ***The 1992 Advanced Placement test results in English and mathematics by score level and as a percent of high school graduates***

AP Score	Math/ Calculus AB	Math/ Calculus BC	Total	Total as a Percent of High School Graduates	Cumulative Percent Scoring 3 or Higher
1	12,855	1,760	14,615	.6	
2	13,077	1,121	14,198	.6	
3	19,647	3,724	23,371	.9	.9
4	15,424	3,472	18,896	.8	1.7
5	13,916	5,318	19,234	.8	2.5
Total	74,919	15,395	90,314	3.7	

AP Score	English Language/ Composition	English Literature/ Composition	Total	Total as a Percent of High School Graduates	Cumulative Percent Scoring 3 or Higher
1	1,551	3,247	4,798	.2	
2	10,431	31,235	41,666	1.7	
3	10,440	42,775	53,215	2.1	2.1
4	5,803	21,179	26,982	1.1	3.2
5	2,367	12,444	14,811	.6	3.8
Total	30,592	110,880	141,472	5.7	

NOTE: Students cannot take both the Math AB and the Math BC examinations in the same year. Therefore, there is no overlap in these counts. The two English test populations overlapped by 167 students in 1992.

SOURCE: Advanced Placement Program of the College Board, *The 1992 Advanced Placement Program National [and State] Summary Reports* (New York, NY: Author, 1992).

achievement compared to 13-year-olds in other nations participating in the International Assessment of Educational Progress (IAEP).[6]

If one of the purposes of the standard-setting effort is to establish expectations whereby the nation's students will be competitive with those in other countries (i.e., will meet "world-class standards"), then it is reasonable to ask to what degree the standards represent achievement levels actually attained by students in those countries. Given mean differences between other countries and the United States, one would expect greater percentages of students in countries such as Korea and Switzerland to perform at the advanced and proficient levels. However, if it turned out that only small percentages of students in other countries reached these NAEP levels, then the levels would appear to have been set unreasonably high even for other nations.

Unfortunately, comparable international and NAEP data sets are available only for eighth-grade mathematics; furthermore, one must rely on imperfect statistical linkages to draw the desired comparisons.

Two separate comparisons are available. The first, by Beaton and Gonzalez, links the 1990 NAEP Trial State Assessment for eighth-grade mathematics with the 1991 IAEP mathematics assessment for 13-year-olds.[7] The second, by Pashley and Phillips, links the 1992 NAEP with the same 1991 IAEP data.[8] Thus, both analyses are intended to draw comparisons between the United States and other nations participating in the 1991 IAEP. Although NAEP and IAEP samples are each probability samples drawn to be representative of the U.S. population, the linkages necessary to translate IAEP results for other countries onto the NAEP scale rest on several assumptions about equivalences that cannot be defended rigorously. Both assessments were based on the same NAEP mathematics framework but differ in reliability; NAEP was administered to eighth graders and IAEP to 13-year-olds (about 58 percent of U.S. eighth graders are 13 years old); the assessments were administered at about the same time of year, February versus March, but in different years; the NAEP Trial State Assessment included only public school students, while the IAEP samples included both public and private schools. Comparability of results was not built into the study design (as it is, for example, with the Third International Mathematics and Science Study [TIMSS] planned for 1994-95 and 1998-99).

[6] These rankings were based on mean scores and cannot be immediately translated into comparisons at advanced-score levels or in other regions of the distribution. We could expect that many more Taiwanese than U.S. students would score at an advanced level because the average IAEP score for Taiwan was 297 compared to the United States at 262. However, it cannot be assumed that differences at the mean translate simply into corresponding differences in the upper (or lower) tails of the distributions. For example, in comparing states in the 1990 Trial State Assessment, the ranking of states changed considerably when they were ranked by their 90th percentile scores instead of their mean scores. (R.L. Linn, L. Shepard, and E. Hartka, "The Relative Standing of States in the 1990 Trial State Assessment: The Influence of Choice of Content, Statistics, and Subpopulation Breakdowns," in *Studies for the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment* [Stanford, CA: The National Academy of Education, 1991].)

[7] A.E. Beaton and E.J. Gonzalez, "Comparing the NAEP Trial State Assessment Results with the IAEP International Results," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993).

[8] P.J. Pashley and G.W. Phillips, *Toward World-Class Standards: A Research Study Linking International and National Assessments* (Princeton, NJ: Educational Testing Service, June 1993).

The two analyses presented here differ in important respects in addition to the year of NAEP TSA data used. Beaton and Gonzalez used linear equating and assumed random equivalence between the U.S. IAEP and NAEP samples. Pashley and Phillips used linear regression to make a projection from one assessment to the other and linked the two using a sample of 1,609 U.S. eighth graders who took both assessments in 1992. The differences in results between the two analyses illustrate the difficulties inherent in making statistical adjustments and in trying to base comparisons on "noncomparable" data. Estimated percentages of the population above the NAEP achievement levels for countries participating in IAEP are shown for the two analyses in tables 4.11 and 4.12, respectively. Note that U.S. results are also shown as "estimates" which means that the same statistical procedures were applied to estimate the performance of U.S. students on NAEP using the U.S. IAEP sample.

The large differences found by the two studies are unfortunate because they lead to quite different conclusions about the reasonableness of the NAEP achievement levels. The Beaton and Gonzalez results show very high percentages of students in Taiwan and Korea scoring at the proficient level or above: 54 percent and 52 percent compared to 17 percent in the United States. For Switzerland and France, the results at the advanced level are much more similar to U.S. results (4 percent compared to U.S. 3 percent), but these countries still substantially exceed the United States in percentages of students at the proficient level. If the results of this analysis were correct, the achievement-level cutpoints for eighth-grade mathematics would not appear to be too high.

The results of the Pashley and Phillips analysis are markedly different, with much smaller percentages of students estimated to be advanced, even in Korea and Taiwan. The connotation of "world-class standards" is that students in other countries are *routinely* performing at these levels. If only 6 percent of students in Korea, and smaller percentages in other countries, are achieving at the advanced level, then the level is rarely attainable even by international standards. The results of the second analysis, if true, in combination with other findings such as the AP data presented above, suggest that the NAEP achievement-level cutpoints have been set too high.

Because the two analyses differ in several respects, it is not a simple matter to adjudicate which is more trustworthy. A simpler, more intuitive check, which does not require any statistical estimation, is to compare U.S. performance with that of other countries on the IAEP alone. In table 4.13, detailed percentile data, available for only the top two countries, Taiwan and Korea, were used to report IAEP results at cutscores corresponding to the U.S. advanced, proficient, and basic achievement-level percentages. These data support the reasonableness of estimates obtained from the Beaton and Gonzalez analysis. High percentages of students scoring above the advanced and proficient cutpoints in two Asian nations suggest that the NAEP achievement levels for eighth-grade mathematics have not been set "too high" in a normative sense, given that students are clearly able to attain these scores. International comparisons cannot, however, be translated into specific achievement levels because choice of countries and percentile points would each lead to different cutpoints. (International comparisons also do not help evaluate the substance of the achievement levels. The achievement levels are "criterion referenced" in the sense that they describe specific performance levels. Therefore, knowing that "more" students in another country can do certain things does not help evaluate what percentage of students in the United States can do those things.)

Table 4.11. ***Percent of population at NAGB achievement levels for the 1990 eighth-grade NAEP Trial State Assessment in mathematics and 1991 IAEP mathematics for 13-year-olds from the Beaton and Gonzalez analysis***

IAEP POPULATION							
TSA State	Mean	(s.e.)	Below Basic	Basic	Proficient	Advanced	Proficient & Advanced
TAIWAN	296.7	(1.5)	19.7	26.1	29.7	24.4	54.1
KOREA	294.1	(1.3)	15.7	32.2	36.5	15.6	52.2
SOVIET UNION	287.6	(1.5)	14.7	40.8	38.5	5.9	44.4
SWITZERLAND	287.5	(1.9)	12.3	45.8	37.6	4.2	41.8
HUNGARY	284.8	(1.4)	18.7	41.1	32.5	7.6	40.1
North Dakota	281.1	(1.2)	19.0	47.3	29.8	4.0	33.8
Montana	280.5	(0.9)	19.6	47.8	28.5	4.1	32.6
FRANCE	278.1	(1.3)	23.3	45.3	27.7	3.7	31.4
Iowa	278.0	(1.1)	23.7	45.9	26.7	3.8	30.4
ISRAEL	276.8	(1.3)	22.2	48.7	27.0	2.1	29.1
ITALY	276.3	(1.4)	24.6	45.2	27.7	2.5	30.1
Nebraska	275.7	(1.0)	25.7	44.4	26.3	3.6	29.9
Minnesota	275.4	(0.9)	26.2	45.2	25.0	3.7	28.7
Wisconsin	274.5	(1.3)	27.7	43.5	25.2	3.6	28.8
CANADA (Global)	274.0	(1.0)	25.4	50.8	21.4	2.3	23.7
New Hampshire	273.2	(0.9)	28.7	46.3	21.8	3.2	25.0
SCOTLAND	272.4	(1.5)	28.3	47.1	22.3	2.2	24.5
Wyoming	272.2	(0.7)	28.9	47.6	21.6	2.0	23.6
Idaho	271.5	(0.8)	29.9	47.1	21.5	1.5	23.0
IRELAND	271.4	(1.4)	29.5	45.3	22.7	2.5	25.2
Oregon	271.4	(1.0)	31.8	43.1	21.6	3.5	25.1
Connecticut	269.9	(1.0)	34.1	39.7	22.4	3.9	26.2
New Jersey	269.7	(1.1)	35.5	39.3	21.3	3.9	25.3
Colorado (TSA)	267.4	(0.9)	35.7	42.8	19.4	2.2	21.6
SLOVENIA	267.3	(1.3)	33.7	48.4	16.4	1.5	17.9
Indiana	267.3	(1.2)	36.8	42.6	17.6	3.0	20.6
Pennsylvania	266.4	(1.6)	37.1	41.5	19.1	2.3	21.4
Michigan	264.4	(1.2)	39.7	40.6	17.4	2.4	19.7
Virginia	264.3	(1.5)	42.3	37.0	16.6	4.1	20.8
COLORADO (IAEP)	264.2	(0.7)	38.5	44.2	15.3	2.0	17.3
Ohio	264.0	(1.0)	40.4	41.1	16.6	2.0	18.5
Oklahoma	263.2	(1.3)	40.8	42.4	15.3	1.6	16.9
SPAIN	261.9	(1.3)	39.0	50.3	10.3	0.4	10.7
UNITED STATES (IAEP)	261.8	(2.0)	42.0	40.7	14.4	2.9	17.3
United States (NWP)	261.8	(1.4)	42.7	38.1	16.9	2.3	19.2
New York	260.8	(1.4)	43.5	37.7	15.7	3.1	18.8
Maryland	260.8	(1.4)	44.2	36.0	16.8	3.1	19.9
Delaware	260.7	(0.9)	45.5	35.9	16.5	2.1	18.6
Illinois	260.6	(1.7)	43.1	38.8	16.2	2.0	18.2
Rhode Island	260.0	(0.6)	45.1	36.6	16.4	1.8	18.3
Arizona	259.6	(1.3)	45.2	38.7	14.6	1.5	16.1
Georgia	258.9	(1.3)	46.6	36.2	14.6	2.6	17.2
Texas	258.2	(1.4)	48.0	36.1	14.0	2.0	15.9
Kentucky	257.1	(1.2)	49.5	37.0	12.3	1.2	13.5
New Mexico	256.4	(0.7)	49.3	37.9	11.6	1.2	12.8
California	256.3	(1.3)	49.1	35.0	13.9	2.0	15.9
Arkansas	256.2	(0.9)	48.8	38.9	11.5	0.9	12.4
West Virginia	255.9	(1.0)	50.7	37.2	11.0	1.1	12.1
Florida	255.3	(1.3)	50.8	34.4	13.1	1.7	14.8
Alabama	252.9	(1.1)	52.8	35.5	10.6	1.1	11.7
Hawaii	251.0	(0.8)	54.7	31.0	12.5	1.8	14.3
North Carolina	250.4	(1.1)	55.6	33.2	10.5	0.8	11.3
Louisiana	246.4	(1.2)	61.5	31.0	7.0	0.6	7.6
JORDAN	236.1	(1.9)	70.5	24.8	4.5	0.1	4.7
Guam	231.8	(0.7)	73.3	21.5	4.7	0.5	5.2
District of Columbia	231.4	(0.9)	79.0	17.2	3.0	0.8	3.8
Virgin Islands	218.7	(0.9)	89.6	9.5	0.9	0.0	0.9

Table 4.11. ***Percent of population at NAGB achievement levels for the 1990 eighth-grade NAEP Trial State Assessment in mathematics and 1991 IAEP mathematics for 13-year-olds from the Beaton and Gonzalez analysis (continued)***

Populations with Exclusions and Low Participation Rates							
CHINA	307.4	(2.1)	4.0	28.1	48.6	19.3	67.9
ENGLAND	272.1	(3.6)	28.4	46.7	21.5	3.4	24.9
PORTUGAL	251.7	(1.6)	51.9	41.0	6.7	0.4	7.1
SAO PAULO (BRAZIL)	228.4	(2.0)	75.5	21.4	3.1	0.1	3.2
FORTALEZA (BRAZIL)	217.8	(1.6)	83.7	15.1	1.2	0.0	1.2
MOZAMBIQUE	214.4	(1.2)	96.1	3.9	0.0	0.0	0.0

SOURCE: A.E. Beaton and E.J. Gonzalez, "Comparing the NAEP Trial State Assessment with the IAEP International Results," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 383-384.

Table 4.12. ***Percentage (confidence) intervals at or above NAGB achievement levels for 1991 IAEP mathematics for comprehensive populations—age 13 public and private schools from the Pashley and Phillips analysis***

Comprehensive Populations	Percentage (Confidence) Intervals of Students At or Above NAGB Achievement Levels		
	Basic (256)	Proficient (294)	Advanced (331)
Korea	78.7-82.5	34.6-39.3	05.3-07.5
Taiwan	75.9-80.1	38.2-43.1	08.8-12.0
Switzerland 15 Cantons	82.7-85.1	31.6-34.6	02.8-03.9
Soviet Union Russian-speaking Schools in 14 Republics	77.9-81.4	28.1-31.9	02.4-03.7
Hungary	74.6-78.7	26.6-30.6	02.6-04.0
France	70.7-74.6	21.4-25.0	01.5-02.5
Emilia-Romagna, Italy	70.2-74.2	20.2-23.9	01.2-02.1
Israel Hebrew-speaking Schools	70.5-74.5	19.8-23.3	01.1-01.9
Canada	68.9-71.6	18.2-20.6	01.1-01.7
Scotland	66.9-71.0	17.8-21.2	01.0-01.8
Ireland	65.5-69.8	17.5-20.9	01.0-01.9
Slovenia	62.3-66.7	14.4-17.5	00.6-01.3
Spain Spanish-speaking Schools except in Cataluña	57.5-61.8	10.4-13.0	00.3-00.7
United States	55.6-60.5	12.6-15.9	00.8-01.7
Jordan	35.4-40.2	04.6-06.5	00.1-00.3

SOURCE: P.J. Pashley and G.W. Phillips, *Toward World-Class Standards: A Research Study Linking International and National Assessments* (Princeton, NJ: Educational Testing Service, June 1993).

Table 4.13. ***1991 IAEP results for Taiwan, Korea, and U.S. based on U.S. percentiles corresponding to advanced, proficient, and basic***

Country	Percent Below Basic	Percent At or Above		
		Basic	Proficient	Advanced
Taiwan	19.3	80.7	58.5	24.4
Korea	14.2	85.8	59.1	15.4
U.S.	39.5	60.5	22.5	3.0

Note: U.S. percentiles are the average of 1990 and 1992 results.

SOURCE: Educational Testing Service, Unpublished data.

International Assessment results are typically reported for various percentiles, as well as for the mean score. These data are shown in table 4.14 for several important comparison countries. Note that the IAEP scores are reported as "percent correct" on the total assessment, with 100 percent obviously the highest score possible. Notice also that students at the 99th percentile for the United States got 97.3 percent correct on the assessment. This value corresponds to the 90th percentile of students in Taiwan, meaning that the top 10 percent of students in Taiwan can do what only the top 1 percent of U.S. students can do. This comparison prompts an important observation. *Meaningful and important international comparisons can be obtained without the use of achievement levels.* Cross-national comparisons such as those in table 4.14 are just as effective as those in table 4.13 in communicating the fact that students in the United States lag seriously behind those of our economic competitors, both at the mean and in the upper tails of the distributions.

In both tables 4.13 and 4.14, the comparisons of U.S. students' performance to that in other countries are made using IAEP data. For benchmarking purposes, it would be possible to use rough percentile equivalences to make the same kinds of comparisons on NAEP. (In fact, percentile equivalences were used to create the international percentages in table 4.13. The U.S. percentage classified as advanced *on NAEP* was 3 percent, so the U.S. 97th percentile *on IAEP* was used as a benchmark to see what percentage of students in other countries were above this same point.)

Kentucky Comparison for Eighth-Grade Mathematics

Although many states have assessment programs that would permit indirect comparisons with NAEP results, Kentucky is rare in that it uses achievement levels similar to the NAEP levels. In a report to the Governing Board, Commissioner Boysen from Kentucky pointed out the close similarities between the four categories on NAEP and the corresponding four categories for the Kentucky Instructional Results Information System (KIRIS). Category labels and brief generic definitions are shown in table 4.15. Although the KIRIS definitions are phrased in terms of task completion

Table 4.14. ***The 1991 IAEP results for selected countries: Percentage of assessment items answered correctly by students at each country's mean and at the 90th, 95th, and 99th percentiles***

Country	Mean	90th	95th	99th
Taiwan	72.7	97.3	98.7	100.0
Korea	73.4	96.0	97.3	100.0
Switzerland	70.8	93.3	94.7	98.7
Soviet Union	70.2	92.0	94.7	98.7
United States	55.3	82.7	90.7	97.3

Note: U.S. percentages are the average of 1990 and 1992 results.

Table 4.15. ***Comparison of NAEP and Kentucky definitions of performance levels***

NAEP	KIRIS
Below Basic. Students have little or no mastery knowledge and skills necessary to perform work at each grade level.	**Novice.** The student is beginning to show an understanding of new information or skills.
Basic. Students have partial mastery of knowledge and skills fundamental for proficient work.	**Apprentice.** The student has gained more understanding, can do some important parts of the task.
Proficient. Students demonstrate competency over challenging subject matter and are well prepared for the next level of schooling.	**Proficient.** The student understands the major concepts, can do almost all of the task, and can communicate concepts clearly.
Advanced. The student shows superior performance beyond the proficient grade-level mastery. Advanced students at the 12th-grade level are ready for rigorous college-prep courses, career preparation, or other training.	**Distinguished.** The student has deep understanding of the concept or process and can complete all important parts of the task. The student can communicate well, think concretely and abstractly, and analyze and interpret data.

SOURCE: Thomas C. Boysen, Commissioner, Kentucky Department of Education, "Importance of State-level NAEP Reports" (Presentation to the National Assessment Governing Board, May 14, 1993).

rather than degree of mastery of a body of knowledge, the implied levels are similar, especially for the upper two categories.

The levels for the Kentucky mathematics assessment were determined on the open-ended performance tasks by asking a content committee to look at sample papers for each score level. Each student responded to three tasks scored 1-4; therefore, possible scores ranged from 0-12. The committee studied samples of student work at 12, 11, and so forth. The 1992 NAEP mathematics assessment was administered to a random sample of Kentucky eighth graders while KIRIS was given to the entire population of Kentucky eighth graders. The results of the two assessments produced similar percentages of students in the upper two categories, as shown in table 4.16. These similarities are surprising given that both the measures and methods of standard setting were so different. Nonetheless, *they support the reasonableness of the NAEP eighth-grade proficient and advanced cutpoints in mathematics because two different assessments arrived independently at similar empirical results for these categories.*

Table 4.16. ***Comparison of Kentucky 1992 results for eighth-grade mathematics on NAEP and KIRIS***

NAEP		KIRIS	
Advanced	2	Distinguished	3
Proficient	15	Proficient	10
Basic	38	Apprentice	23
Below Basic	43	Novice	65

SOURCE: Thomas C. Boysen, Commissioner, Kentucky Department of Education, "Importance of State-level NAEP Reports" (Presentation to the National Assessment Governing Board, May 14, 1993).

Grade-to-Grade Fluctuations Using the Achievement-Level Classifications

In early evaluations of the 1990 achievement levels in mathematics, grade-to-grade "coherence" was established as a minimal criterion for judging the reasonableness of the levels. Higher levels of achievement should be expected in higher grades. When levels were set such that students at the 8th-grade basic level were expected to outperform 12th-grade basic students on common items, the levels were said to lack coherence. As described in chapter 3, grade-to-grade coherence in the final 1992 levels was virtually assured by the graphs used to display grade-to-grade patterns and the feedback given to panelists at the end of each round of judgments. Indeed, the cutpoints set in both reading and mathematics have the desired property of increasing stringency by grade level.

A similar issue regarding the reasonableness of the achievement levels has to do with their relative difficulty *within* grade. Percentages of students at or above each achievement level are reported in tables 4.17 and 4.18. What kinds of interpretations are likely to be drawn from the finding that in reading 4.5 percent of fourth graders are classified as advanced but only 2.1 percent of eighth graders are so classified? More seriously, what does it mean in mathematics that 25 percent of eighth graders are classified as proficient or above but only 16 percent of 12th graders? Are high school teachers doing a poorer job of instruction than middle school teachers? Is it that a significant proportion of 12th graders have had little or no mathematics since the 8th grade, and, therefore, actually perform less well than 8th graders? Or are these apparent grade-to-grade differences an artifact of the inaccuracy of standard setting at each grade level? The last is the most likely explanation, especially when the fluctuations are erratic from grade to grade. Perhaps the increasing percentages at basic and above in reading from grade 4 to grade 12 can be explained by the fact that the school population becomes more select as students drop out from grade to grade. But this would not explain why there are fewer advanced readers in grade 8 than grade 12. And if the population is becoming more select, why are the basic percentages nearly uniform in mathematics? *Given that cutpoints are arbitrary and fallible, NAEP reports might need to carry warnings that grade-to-grade interpretations of the kind suggested above are not warranted.* In the future, the reasonableness of grade-to-grade patterns might be considered as one factor, along with sources of external comparison data and the like, in deliberations about where cutpoints should be drawn.

Content-Expert Evaluations

Thus far in this chapter, the Panel has considered only external empirical comparisons. That is, we have been answering the question, "What do performance data from various external referent groups tell us about the reasonableness of the NAEP achievement levels?" In addition, the Panel was interested in the appropriateness of the levels from the perspective of content experts. The purpose of achievement levels is to report NAEP results in terms of what students *should* be able to do. Do reading and mathematics experts agree with the expectations represented by the levels? Although content experts should not be the sole authority in setting academic standards, the usefulness of the levels would be seriously jeopardized if experts found them to be inaccurate or inadequate.

For the purposes of these studies, "content experts" who have been leaders in their fields were used. They were expected to have extensive knowledge of curriculum and instruction theories in their respective disciplines in addition to classroom teaching experience. Experts in reading and mathematics had to be familiar with common conceptions and expectations supported by the profession, as well as recognize areas of controversy. For example, all experts in mathematics were familiar with the NCTM Standards. Content experts were also expected to be familiar with NAEP and with the NAEP framework in their subject area. These requirements for expertise are more like the criteria used by NAGB in constituting expert panels to develop content frameworks than the criteria used to select teachers and nonteacher educators for the initial achievement-level-setting process. NAGB included some experts of this type for their second-stage validation meetings.

Table 4.17. ***The 1990 and 1992 national overall average mathematics proficiency and percentage of students at or above achievement levels, grades 4, 8, and 12***

Grades	Assessment Years	Average Proficiency	Percentage of Students At or Above			Percentage Below Basic
			Advanced	Proficient	Basic	
4	1992	218(0.7)>	2(0.3)	18(1.0)>	61(1.0)>	39(1.0)<
	1990	213(0.9)	1(0.4)	13(1.1)	54(1.4)	46(1.4)
8	1992	268(0.9)>	4(0.4)	25(1.0)>	63(1.1)>	37(1.1)<
	1990	263(1.3)	2(0.4)	20(1.1)	58(1.4)	42(1.4)
12	1992	299(0.9)>	2(0.3)	16(0.9)	64(1.2)>	36(1.2)<
	1990	294(1.1)	2(0.3)	13(1.0)	59(1.5)	41(1.5)

> The value for 1992 was significantly higher than the value for 1990 at about the 95 percent confidence level. < The value for 1992 was significantly lower than the value for 1990 at about the 95 percent confidence level. The standard errors of the estimated percentages and proficiencies appear in parentheses. It can be said with 95 percent confidence that for each population of interest, the value for the whole population is within plus or minus two standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference.

SOURCE: I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *NAEP 1992 Mathematics Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, 1993), 4.

Table 4.18. ***The 1992 percentage of students at or above reading achievement levels***

Grades	Percentage At or Above			Percentage Below Basic
	Advanced	Proficient	Basic	
4	4.5	25.3	59.0	41.0
8	2.1	27.5	68.8	31.2
12	3.2	37.0	75.2	24.8

SOURCE: Educational Testing Service, *NAEP 1992 Reading Report Card for the Nation and the States* (Princeton, NJ: Author, forthcoming).

The content-expert studies were directed by independent teams of researchers in mathematics and in reading. In each case, panels of experts were convened (14 experts in mathematics, 17 experts in reading) and asked to evaluate (1) the appropriateness of the level cutpoints, and (2) the quality of the descriptions. Reports of findings are based on quantitative results as well as field notes and transcripts of meetings.[9]

To facilitate their evaluation of the level cutpoints, experts were shown the NAEP continuum and were given detailed information about the kinds of items students could do at various points along the score scale. Items were "mapped" onto the score scale at the point where 80 percent of students in that grade could answer the item correctly (after allowing for guessing). A similar item-mapping procedure was used to report results of the NAEP adult literacy study.[10] It is a useful technique for showing the kinds of things that students at each score point can do. For extended open-ended NAEP items scored 0-4, sample responses for each score were mapped onto the scale. Thus a "2" response to an item might fall in the basic region, while a "4" response to the same item could be in the proficient or advanced region. Given this substantive picture of a well-described performance continuum, it was then possible to ask where the line should be drawn between basic and proficient and so forth.

In both subject areas, experts were asked to apply the final version of the NAEP descriptions in making their judgments about cutpoints because the final descriptions would be used to report the meaning of each level to the public. Experts were identified by grade level and were further subdivided into two teams per grade. The grade-level subgroups first worked independently using randomly equivalent item maps (where each group was looking at a random half of the items) and then came together to reconcile differences. The groups were not forced to agree on specific cutpoints but were asked to identify ranges, "gray areas" or "regions of uncertainty," where cutpoints could plausibly be located. The question was not whether experts would come to precisely the same cutpoints as developed by NAGB but whether the NAGB cutpoints were within ranges considered to be defensible by experts.

Results for mathematics. In mathematics, experts used item maps to identify reasonable ranges for cutscores separately for each content strand. Results are shown for grades 4, 8, and 12 in tables 4.19, 4.20, and 4.21. In some cases, it was possible for the teams to come together and agree on a very specific point distinguishing two levels. This would occur when items on either side of a point represented a clear distinction, with items below the point fitting the lower description and items above that point fitting the higher description. In many cases, however, panelists agreed on fairly large "gray areas" with the understanding that a reasonable cutpoint could be anywhere within that region. This occurred primarily because of ambiguity in trying to apply the descriptions or sometimes because there were gaps with no items to use to make the distinction. In extreme cases, there were no items at all matching a level (or the implied below basic level), and no cutpoint could be set.

[9] E.A. Silver and P.A. Kenney, "Expert Panel Review of the 1992 NAEP Mathematics Achievement Levels," and D. Pearson and L. DeStefano, "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report Three: Comparison of Cutpoints for the 1992 NAEP Reading Achievement Levels with Those Set by Alternate Means," both in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993).

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

Table 4.19. ***Grade 4 expert panel cutpoint decisions by content strand***

Content Strand	Below Basic/Basic	Basic/Proficient	Proficient/Advanced
Geometry (n = 27 items)	210 - 220	235	295 - 310
Numbers and Operations (n = 64 items)	150 - 155	230 - 250	310 - 335
Measurement (n = 29 items)	231 - 235	255	330
Algebra and Functions (n = 14 items)	195 - 220	250 - 260	290 - 300
Data Analysis, Statistics, and Probability (n = 20 items)	220	250 - 258	330
Estimation* (n = 13 items)	none	220 - 240	335 - 345

* Estimation items were not included in the composite.

SOURCE: E.A. Silver and P.A. Kenney, "Expert Panel Review of the 1992 NAEP Mathematics Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 223.

Table 4.20. ***Grade 8 expert panel cutpoint decisions by content strand***

Content Strand	Below Basic/Basic	Basic/Proficient	Proficient/Advanced
Geometry (n = 36 items)	221 - 236	296 - 302	408 - 482
Numbers and Operations (n = 58 items)	none	336 - 337	356 - 370
Measurement (n = 32 items)	238 - 243	312 - 330	412 - 415
Algebra and Functions (n = 29 items)	206-261	309 - 311	352
Data Analysis, Statistics, and Probability (n = 28 items)	none	none	378 - 392
Estimation* (n = 20 items)	none	390 - 442	none

* Estimation items were not included in the composite.

SOURCE: E.A. Silver and P.A. Kenney, "Expert Panel Review of the 1992 NAEP Mathematics Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 226.

Table 4.21. ***Grade 12 expert panel cutpoint decisions by content strand***

Content Strand	Below Basic/Basic	Basic/Proficient	Proficient/Advanced
Geometry (n = 31 items)	285 - 295	330 - 335	372
Numbers and Operations (n = 43 items)	268 - 271	341 - 343	375 - 385
Measurement (n = 28 items)	258 - 273	331 - 333	none
Algebra and Functions (n = 47 items)	215 - 272	321 - 322	358 - 362
Data Analysis, Statistics, and Probability (n = 29 items)	none	311 - 316	407 - 489
Estimation* (n = 19 items)	none	349 - 367	none

* Estimation items were not included in the composite.

SOURCE: E.A. Silver and P.A. Kenney, "Expert Panel Review of the 1992 NAEP Mathematics Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 231.

Because NAEP achievement levels are reported only for the composite mathematics scale, content experts in mathematics next had to convert their cutscores for the separate strands into cutpoints on the composite scale. They were asked to consider explicitly what the most defensible method would be for combining results across strands. Mathematics experts at each grade level argued against the use of a composite score on substantive grounds. At grade 4, for example, the heavy representation of the numbers and operations content strand (45 percent) would obscure the contribution of other strands. At grade 8, lack of opportunity to learn content from the algebra and functions strand (weighted 20 percent) could distort the picture of student proficiency. Ultimately, however, the panelists agreed to satisfy the study demands for composite cutpoints by computing weighted averages of both the upper and lower boundaries of each cutpoint based on the weights assigned to each strand in the framework.[11] The composite cutpoint ranges recommended by the mathematics experts are shown in table 4.22.

Table 4.22. ***Comparison of expert panel composite cutpoint intervals with official NAGB cutpoints set in mathematics***

	Below Basic/Basic	Basic/Proficient	Proficient/Advanced
		GRADE 4	
EP Cutpoint Interval	187 - 194	240 - 250	312 - 326
Official (NAGB) Cutpoint	211	248	280
		GRADE 8	
EP Cutpoint Interval	216 - 230*	303 - 312	377 - 404
Official (NAGB) Cutpoint	256	294	331
		GRADE 12	
EP Cutpoint Interval	255 - 274	328 - 331	382 - 398
Official (NAGB) Cutpoint	287	334	366

* This composite cutpoint interval was based on an estimate for the cutpoint interval for the data analysis, statistics, and probability content strand. The grade 8 panelists agreed that, for this strand, the 1992 eighth-grade assessment contained no items that were representative of below basic or basic performance; therefore, they could not identify the interval that contained the cutpoint. For the purposes of determining a composite cutpoint interval, the scale value 215 (i.e., the midpoint between 200 and 230 [the scale value for the first "proficient" item in a content strand]) was used in the calculations.

SOURCE: E.A. Silver and P.A. Kenney, "Expert Panel Review of the 1992 NAEP Mathematics Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 235.

[11] Estimation was not included in the NAEP composite because estimation items were still considered to be experimental in the 1992 assessment.

Across the three grades, mathematics content experts consistently set cutpoints for the basic level below the official NAEP achievement-level cutpoints and above the official cutpoints for the advanced level. The official NAEP proficient cutpoint was within the interval identified by experts at grade 4 and relatively close to it for grades 8 and 12. Based on records kept of explanations given during the working sessions, these patterns of higher and lower cutpoints were clearly linked to correspondences between specific item features and phrases in the level descriptions. Content experts moved the basic cutpoint down the score scale (compared to the official cutoffs) because they judged items lower on the score scale as consistent with the basic description. Experts made these decisions in deference to the specific content of the descriptions even when they judged some items to be "below basic" in terms of their own personal definitions of basic performance. In contrast, at the upper end of the score scale, faithful adherence to the achievement-level descriptions meant that experts were unwilling to accept difficult, but conceptually simple items as evidence of advanced performance, because such items did not match the narrative descriptions. They therefore reached higher and higher on the score continuum in search of items consistent with the descriptions.

Results for reading. The results of expert-determined cutpoints in reading are presented in table 4.23 in comparison to the official reading cutpoints. *With the exception of the grade 12 basic and proficient categories, content experts in reading consistently identified cutpoints above the official achievement-level cutpoints.* The explanation for setting higher standards was very similar to what occurred at the advanced level in mathematics for all three grades. Experts were forced to move up the NAEP scale to find items that matched specific elements in the achievement-level descriptions.

In the reading study, experts were asked explicitly to compare the cutpoints they set with the official NAEP versions. Experts argued against the official cutpoints by pointing to inconsistencies between the achievement-level descriptions and items mapped near the cutpoints. For example, pointing to 244 for eighth-grade basic, a panelist noted that "most of the items deal with literal comprehension. You have to go up the scale a ways before you get to items that even begin to address the most simple interpretations, inferences, and authors' intent. There is nothing that even deals with making predictions."[12]

The grade 12 basic and proficient results must be treated as anomalous on grounds other than their departure from the above trend. One of the two 12th-grade teams used their own definitions of basic, proficient, and advanced rather than applying the official NAGB descriptions to define achievement. These experts thought the descriptions of basic and proficient were unrealistically high, given their experiences with high school students.[13] Therefore, their results could not be used as examples of cutpoints that follow from a substantive analysis of the descriptions.

[12] Pearson and DeStefano, op. cit., 406.

[13] As an aside, the Panel notes that the term "basic" occasionally raises problems of the kind that occurred with one 12th-grade team because of its connotation of minimal competence. Although this issue was not systematically investigated as part of the Panel's evaluation, NAGB should be aware of problems with the "basic" label. A similar reaction may explain the decision of the National Goals Panel to ignore the basic designation and report only proficient and above scores as "competent" performance. The use of other terms, such as "novice" or "apprentice" adopted in the Kentucky assessment, might be more in keeping with NAGB's intended meaning for the lowest achievement level. However, each set of terms invites inferences that may or may not be supportable by validity evidence. Therefore, the choice of labels and what they are taken to mean by educators and the public should be investigated further.

Table 4.23. ***Comparison of Expert Panel range and final cutpoints with official NAGB cutpoints set in reading***

	Below Basic/Basic	Basic/Proficient	Proficient/Advanced
		GRADE 4	
Expert Panel	225	255	289
Range (2 teams)	221 - 236	245 - 263	285 - 295
Official NAGB Cutpoint	212	243	275
		GRADE 8	
Expert Panel	258	293	350
Range (2 teams)	258 - 259	287 - 293	348 - 355
Official NAGB Cutpoint	244	283	328
		GRADE 12	
Expert Panel	255*	295*	380*
Range (2 teams)	250* - 270	272* - 310	375 - 400
Official NAGB Cutpoint	269	304	348

* One grade 12 team relied on personal definitions of basic, proficient, and advanced rather than the NAEP descriptions.

SOURCE: D. Pearson and L. DeStefano, "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report Three: Comparison of Cutpoints for the 1992 NAEP Reading Achievement Levels with Those Set by Alternate Means," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 405.

Methodological insights from the content-expert studies. Although the content-expert studies were intended to provide yet another perspective on the reasonableness of the final NAEP levels, the Panel did not in fact take the final results as either confirming or disconfirming of the NAGB levels. Instead, in the course of these studies, a great deal more was learned about the process of judgment, features of the NAEP item pools in reading and mathematics, and artifacts introduced by judgmental procedures, including the "procedure" used by the Panel's experts. Previously, researchers in the technical literature on standard-setting methods have found that "different methods lead to different results." However, there has been little understanding about what particular methodological choices might lead to systematically higher or lower standards. Regularities in the content-expert studies, along with internal analyses of data from the St. Louis panelists, suggest that some general lessons can be drawn pertinent to future level-setting efforts.

The approach used in the expert studies was not intended as a new or alternative method. The purpose of providing item maps was to make detailed item information accessible to the experts in a way that illuminated the score scale and made the meaning of the cutpoints directly apparent. However, by documenting the strengths and weaknesses of this "procedure," the Panel can contribute to an understanding of process effects.

There were several positive features of the item-mapping approach, especially when contrasted to the Angoff procedure. These observations are based on comments from the participants and the research directors, who, in the case of reading, had also directed the evaluation study of the Angoff level-setting procedure. First, by design, *the experts or judges using the item-mapping approach had a much more direct understanding of the continuum for which they were attempting to devise levels.* Second, by engaging in discussions and studying the item maps, *participants had a more systematic understanding of the item pool as a whole* than did participants using the Angoff approach. Experts using the item-mapping approach could also clearly say what was missing vis-a-vis the descriptions and identify gaps on the performance continuum. Finally, because content experts in the item-mapping studies were asked directly to produce cutoffs or plausible ranges for cutoffs, their *discussions fostered the development of consensus in all but one instance,* which occurred in reading at grade 12.

As stated previously, consensus does not ensure the validity of standards. However, if consensus is valued, then participants must have direct access to the decision they are trying to make and opportunities to discuss sources of agreement and disagreement. In the Angoff approach, panelists are shown the average of their item ratings but can change the implied cutoff only indirectly by altering individual item p-values; they do not have the opportunity to see items at the average cutpoint and to discuss as a group whether they are satisfied with the standard they have set.

The item-mapping approach also had drawbacks which may have contributed to the tendency for experts to set very high cutoffs in reading for all levels, and in mathematics at the advanced level. Items were mapped at the 80-percent-correct level after correction for guessing. The 80-percent level was selected because it had been used in a previous NAEP application, but for descriptive reporting purposes, not for level setting. A 65-percent mapping criterion could also have been defended as being consistent with the criteria used to select anchor items for the traditional NAEP scale-anchoring procedure discussed in chapter 2. If a lower percent correct value had been selected, items would have been located lower on the scale, probably leading to lower cutpoints for the same item content.

In retrospect, the methodological decision to use a very high mapping criterion of 80-percent-correct appears to have fostered the same kind of presumption that occurred in the implementation of the 1992 achievement-level-setting process with the use of boundary exemplar papers. In that case, as discussed in chapter 3, judges' expectations based on generalizations from sample papers led to cutpoints that were 4 to 8 years in advance of multiple-choice cutpoints. Apparently, when items or sample papers are located on the NAEP scale and used to draw a cutpoint, judges are encouraged to imagine that students at that score point can always do items of that type. However, contrary to the implicit assumption of uniform performance or perfect reliability, items are *imperfectly* correlated with each other and a scale is *not* a perfect hierarchy. Thus, students at a given score level will sometimes get harder items right and miss easier ones.

As we have seen from different study approaches, *judges applying the same definitions are likely to set lower standards when provided with integrated and complete samples of student work than when judging items.* We can speculate that this occurs because judges do not require students to get every "proficient" item right to be considered proficient. Although in theory the estimation of less-than-100-percent-correct p-values should also allow for students to miss some items at each level, item-by-item judgments clearly do not lead to the same results as judgments of complete student work. In this respect, *the item-mapping approach is just as deficient as the Angoff procedure because neither method allows experts or judges to consider what combinations of items and tasks are required to meet the definitions.*

Although mapping at the 80-percent level is one likely explanation for setting levels that appear to be unreasonably high, the lack of adequate advanced items was also mentioned repeatedly during the process as a reason for rejecting cutoffs in the range suggested by other studies. Limitations of the item pools for setting standards and for reporting achievement levels to the public are addressed in more detail in the next section.

The Adequacy of Level Descriptions and Exemplar Items to Represent Content Standards

The division of this chapter into two segments, one evaluating the cutpoints and one addressing the descriptions and sample items, corresponds to two different purposes of the achievement levels. In the first case, achievement levels are intended to answer the "how much" question. How much should students know to be considered advanced, proficient, or basic? This question was addressed in the previous section. In the second case, the purpose of the achievement levels is to convey subject-matter standards to the public and to report what it is substantively that students must be able to do to be considered to be performing at a given achievement level. We now turn to the content of the achievement-level descriptions and exemplar items.

The Panel commissioned groups of subject-matter experts in mathematics and reading to conduct the necessary content analyses. In mathematics, the content-expert group that participated in the item-mapping study formally addressed a set of questions regarding the descriptions and exemplar items. Data were also available from a content validity study conducted as part of the larger evaluation of the Trial State Assessment. In reading, responses were provided by two different groups of experts: the content-expert group convened to do the item-mapping study and a group of nationally recognized, university-based researchers. In addition, information was available from interviews with teachers who participated in the contrasting-groups studies described in the first part of this chapter. In both reading and mathematics, the item-mapping exercise itself also produced systematic comparisons between the level descriptions and NAEP item pools. Findings are reported separately for mathematics and then for reading. A final section is devoted to problems of "should versus can" interpretations of achievement-level results and to the problem of statistically misfitting exemplar items.

After completing their analyses of plausible score ranges for achievement-level cutpoints on the NAEP scale, the content-expert panel in mathematics was asked to address a series of questions about the descriptions they had been applying. For example: (1) *Do the achievement levels reflect professionally defensible expectations for student performance in mathematics at each grade level?* and (2) *How well do the mathematics achievement-level descriptions communicate important outcomes to knowledgeable professionals in the field of mathematics education and to the public?* Because of their participation in the item-mapping study, panelists had extensive familiarity with the entire secure item pool.

The mathematics experts were in favor of establishing achievement levels to use in reporting NAEP results. However, "there was general agreement that the 1992 NAEP mathematics achievement levels may not be completely defensible because of their close ties to the 1992 NAEP Mathematics Framework and items."[14] Why was this finding reported as problematic? Shouldn't the level descriptions be closely tied to the assessment framework? More detailed analysis provided two distinct answers. First, there is concern on the part of mathematics experts that the 1990/1992 NAEP Mathematics Framework, though an improvement over earlier versions, is not congruent with the NCTM content standards. Second, there are more particular concerns that some aspects of the NAEP descriptions are incomplete or are tied in narrow ways to specific features of the current item pool.

The mathematics experts compared the St. Louis and final versions of the descriptions and concluded that they differed in nontrivial ways. The St. Louis version more closely matched the 1992 NAEP framework and items, and the final version better represented contemporary thinking and professional content standards. The panelists much preferred the final version as a statement of achievement aspirations for students. "The final version (with the notable exception of the grade 4 descriptions) emphasizes thinking, reasoning, problem solving, application, and communication; and it describes mathematical performance in ways that seem more compatible with current views of mathematical proficiency. In contrast, the St. Louis version, especially at grade 12, describes mathematics in ways that seem more related to discrete outcomes associated with formal course experiences than to the quality of students' mathematical thinking."[15] Ironically, however, the "improvement" of the final descriptions represented a departure from the framework and item pool, thus creating gaps and making it difficult to classify items into achievement-level categories.

The grade 4 mathematics descriptions, even in the final version, were judged to be insupportable by the panelists because of their emphasis on procedural knowledge. "The grade 4 descriptions, especially for basic performance, seem to imply that students should be operating at the 'knowledge-comprehension-skill' level, and not necessarily engaging in critical thinking."[16] In addition, the fourth-grade descriptions only refer to the five content strands generically. There is no mention of what students

[14] Silver and Kenney, op. cit., 236.

[15] Ibid., 243-244.

[16] Ibid., 237.

should be able to do in measurement, geometry, or data analysis. Gaps of this kind were noted at other grade levels as well. A final limitation of the descriptions reported at all three grade levels was that, due to hold-over language from the earlier version of the descriptions, the descriptions were still closely tied to features of the 1992 item pool. For example, the fourth-grade descriptions refer to use of four-function calculators, rulers, and geometric shapes, which experts saw as narrowly tied to the current item pool.

The most critical finding from the expert analysis was the mismatch between the final descriptions and the NAEP mathematics item pool. The descriptions (except for grade 4) are consistent with the NCTM Standards; they are written in a way that emphasizes key themes of communicating mathematically, making connections, and reasoning mathematically. However, in making the descriptions reflect professional content standards (i.e., the NCTM Standards), they no longer reflect the NAEP item pool. In the judgment of experts, the 1992 item pool, in particular the set of released items, has deficiencies as a measure of the NCTM Standards or the descriptions.

This finding leads to two negative conclusions: *(1) The item pool and (with rare exceptions) exemplar items are inadequate for representing content standards to the public, and (2) given that students took the current assessment and not an assessment aligned to the descriptions, students acquiring certain scores may not in fact be able to do what the "descriptions" describe.* Or, particularly at the advanced level, they might not have had the opportunity to show that they can do what is described.

Deficiencies in the item pool were documented as follows. The most pervasive problem was lack of advanced items for all three grade levels. (A similar difficulty occurred with the 1990 levels.) Panelists noted that this problem was not likely to be resolved "until the achievement-level descriptions are written first (instead of post hoc) and then become the basis for framework development, item writing and test development."[17,18] Another systematic problem was the lack of items to represent achievement levels within content strands. The most extreme instances of missing items can be seen in the tables from the item-mapping study (tables 4.19-4.21), where cutpoints could not be set because of the lack of items. At eighth grade, 3 of the 15 levels could not be set. In addition, missing items affected many more cells in the content matrix, leading to very wide score ranges where distinctions could not be made. For example, the "cutpoint" between below basic and basic for eighth-grade algebra and functions was a score range of more than 50 points because of gaps in the item pool.

Experts also criticized the quality of items, finding that as a whole they did not adequately reflect some of the NCTM content standards. For example, at grade 8 there were no items that required making connections between topics despite the expectation from the descriptions that such abilities would be assessed. At grade 12, panelists noted that available items measured algebraic skills in very technical and limited ways. In particular, it was the judgment of the panelists that the format used for the algebra and functions items (e.g., using variables as exponents) made them more difficult than was desirable to assess the intended concepts and skills.[19]

[17] Ibid., 237.

[18] Preliminary descriptions have now been developed as part of the framework development for geography and history.

[19] Silver and Kenney, op. cit., 232.

Somewhat more positive findings regarding the content of the 1992 mathematics assessment were obtained by a separate expert panel convened to assess the content validity of the fourth-grade item pool.[20] In addition to a study of the match between the item pool and the 1992 framework, these experts systematically classified the grade 4 TSA items with respect to four NCTM themes: problem solving, reasoning, communication, and connections. They found that almost 60 percent of the items matched at least one major theme from the NCTM Standards. However, the content-validity expert panel also concluded that only 14 of 158 fourth-grade items (9 percent) could be considered exemplary. Therefore, despite the closer alignment of the 1992 assessment with the standards, most of the items that "fit" the standards do so in less than perfect ways. This finding helps to explain some of the apparent inconsistencies between the conclusions of the two expert panels. In fact, they both agree that only a small proportion of NAEP items reflect the standards in ways that would be useful to communicate those standards to the public.

The group of experts convened to evaluate the achievement levels had additional concerns about the specific exemplar items selected to report the 1992 results. *After reviewing the exemplar items, the experts concluded that "most were inappropriate, uninteresting, and not at all exemplary. Because some exemplars overemphasized highly procedural aspects of mathematics (i.e., decontextualized symbolic manipulation) rather than mathematical problem solving and reasoning, the panelists commented that featuring such items as exemplary might do more harm than good in communicating to the public appropriate goals for school mathematics."*[21, 22]

To illustrate the kinds of concerns raised about the quality of exemplar items for communicating curricular goals in mathematics, pairs of good and not-so-ideal exemplar items are presented for each of grades 4, 8, and 12 in figure 4.3. All examples were used to exemplify advanced performance in the 1992 report of NAEP mathematics results.[23] The fourth-grade "pockets" problem is a favorite with mathematics experts because it requires that children be able to reason mathematically and to communicate their thinking. In contrast, the not-so-ideal exemplar is a routine application of number sentences learned in school. It focuses on the representational system and does not really permit some other way of figuring out the total number of newspapers. If members of the public had only a procedural and rule-oriented view of mathematics, the not-so-ideal example would reinforce, rather than enlighten, that viewpoint.

At grade 8, the preferred exemplar again relies on the extended-response format. The not-so-ideal example is at the eighth-grade level of difficulty only because the functional relationship in the table is somewhat difficult (e.g., 2n + 1). Functional tables of this sort are now being introduced in third grade. As an eighth-grade "advanced" item, it is less than ideal because students do not have to give the rule or explain their answer. The good example for grade 12 illustrates that challenging reasoning problems can be assessed in a multiple-choice format. In contrast, the not-so-ideal 12th-grade advanced problem tests primarily whether students understand the

[20] E.A. Silver and P.A. Kenney, "The Content and Curricular Validity of the 1992 NAEP TSA in Mathematics" (Stanford, CA: The National Academy of Education, forthcoming).

[21] Silver and Kenney, Expert Panel Review.

[22] Emphasis added.

[23] I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *NAEP 1992 Mathematics Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, 1993), 4.

notation for absolute value. An additional example is provided in figure 4.4 of a problem set for fourth graders from the assessment prototypes developed by the Mathematical Sciences Education Board and the National Research Council. Like the pockets problem and other "best" examples from NAEP, assessment items of this type not only serve to measure what students know but also do a good job of communicating the meaning of content standards.

The expert panelists in mathematics also identified serious statistical problems with the exemplars which are discussed in more detail in the final "should versus can" section of this chapter. Of the 23 exemplar items chosen by NAGB for 1992 reporting that were not special-item types, only 14 had scale scores (using the official cutpoints) in the achievement-level categories for which they were exemplars. An illustration of one such item is shown in figure 4.5, which also appeared in the published report for 1992. The item is presented as an example of basic performance, but in fact only 37 percent of students classified at the basic level could answer it correctly. In other words, the majority of students who have met the basic standard cannot do the item. *This means that in addition to their substantive limitations, some of the mathematics exemplars cannot be defended statistically as indicators of what performance at a given level looks like.* The item in figure 4.5 also illustrates the complaint from experts that basic-level knowledge is often portrayed as low-level and mechanical without an expectation that basic-level students should think.

Level Descriptions and Item Content in Reading

Three different data sources were available to evaluate the achievement-level descriptions in reading: transcripts from the expert panel that participated in the item-mapping study, transcripts from the analysis and discussion by a group of reading researchers, and teacher-interview data collected as part of the field-based contrasting-groups study. The expert groups were asked specifically to address consistency between the descriptions and the conceptualization of reading presented in the framework, consistency across achievement levels, consistency across grades, and conceptual differences among the three versions of the reading descriptions.

The evolution of changes in the reading descriptions closely paralleled findings from mathematics. As documented in chapter 3, St. Louis participants reported that they drew heavily on their own personal experience and opinions to develop the descriptions. Therefore, it is not surprising that the group of reading researchers found "no clear evidence that any aspect of the framework (i.e., definition of a good reader, the three purposes for reading, and the cognitive aspects of reading) was used as the main structure of the St. Louis descriptors."[24] Expert panelists objected to the questions about consistency across grades and achievement levels, saying that the implied hierarchy of "skills" was antithetical to the conception of reading advanced in the framework. Rather than differentiating by skills, the framework reflects a set of reading processes "common to all readers" with the provision that characteristics of the text and purposes for reading would vary with grade and performance level.

[24] D. Pearson and L. DeStefano, "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report Two: An Analysis of the Achievement-Level Descriptions," in *Setting Performance Standards for Student Achievement: Background Studies* (Stanford, CA: The National Academy of Education, 1993), 189.

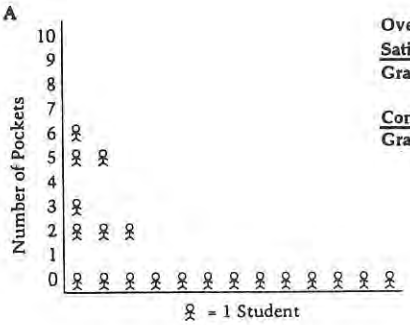
Figure 4.3. ***Examples of exemplary and not-so-ideal exemplar items used to illustrate advanced performance in the 1992 mathematics report***

Good Exemplar Items

Grade 4 Advanced:

There are 20 students in Mr. Pang's class. On Tuesday most of the students in the class said they had pockets in the clothes they were wearing.

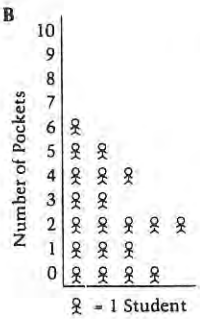
A



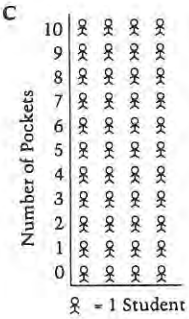
Overall Percent Satisfactory or Better*
Grade 4:10 (0.9)

Conditional-Advanced
Grade 4: 59%

B



C



Which of the graphs most likely shows the number of pockets that each child had? B

Explain why you chose that graph.

Because it shows 20 students and most of the students have pockets.

Explain why you did not choose the other graphs.

It cannot be A because in A most of the students do not have pockets.

It cannot be C because in C there are more than 20 students shown.

Not-So-Ideal Exemplar Items

If ☐ represents the number of newspapers that Lee delivers each day, which of the following represents the total number of newspapers that Lee delivers in 5 days?

A 5 + ☐

(B) 5 × ☐

C ☐ ÷ 5

D (☐ + ☐) × 5

Overall Percent Correct*
Grade 4: 48 (1.2)

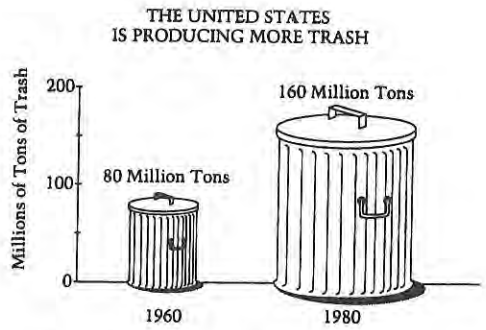
Conditional-Advanced
Grade 4: 95%

*The standard errors of the estimated percentages appear in parentheses.

Figure 4.3. ***Continued***

Good Exemplar Items

Grade 8 Advanced:



The pictograph shown above is misleading. Explain why.

Answer: Both the width and the height have been doubled.

Overall Percent Correct*
Grade 8: 8 (0.8)

Conditional-Advanced
Grade 8: 42

Grade 12 Advanced:

Suppose $4r = 3s = 10t$, where r, s, and t are positive integers. What is the sum of the least values of r, s, and t for which this equality is true?

A 7

B 17

(C) 41

D 82

E 120

Did you use the calculator on this question?

Yes (No)

Overall Percent Correct*
Grade 12: 30 (1.6)

Conditional-Advanced
Grade 12: 84%

*The standard errors of the estimated percentages appear in parentheses.

Not-So-Ideal Exemplar Items

Grade 8 Advanced:

A	B
2	5
4	9
6	13
8	17

14	?

If the pattern shown in the table were continued, what number would appear in the box at the bottom of column B next to 14 ?

A 19

B 21

C 23

D 25

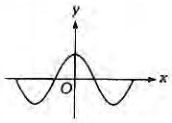
(E) 29

Overall Percent Correct*
Grade 8: 25 (1.3)

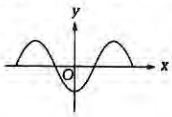
Conditional-Advanced
Grade 8: 79%

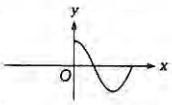
*The standard errors of the estimated percentages appear in parentheses.

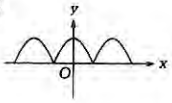
Grade 12 Advanced:

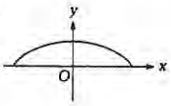


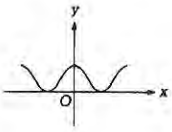
The figure above shows the graph of $y = f(x)$. Which of the following could be the graph of $y = |f(x)|$?

A


B


(C)


D


E


Overall Percent Correct*
Grade 12: 20 (1.3)

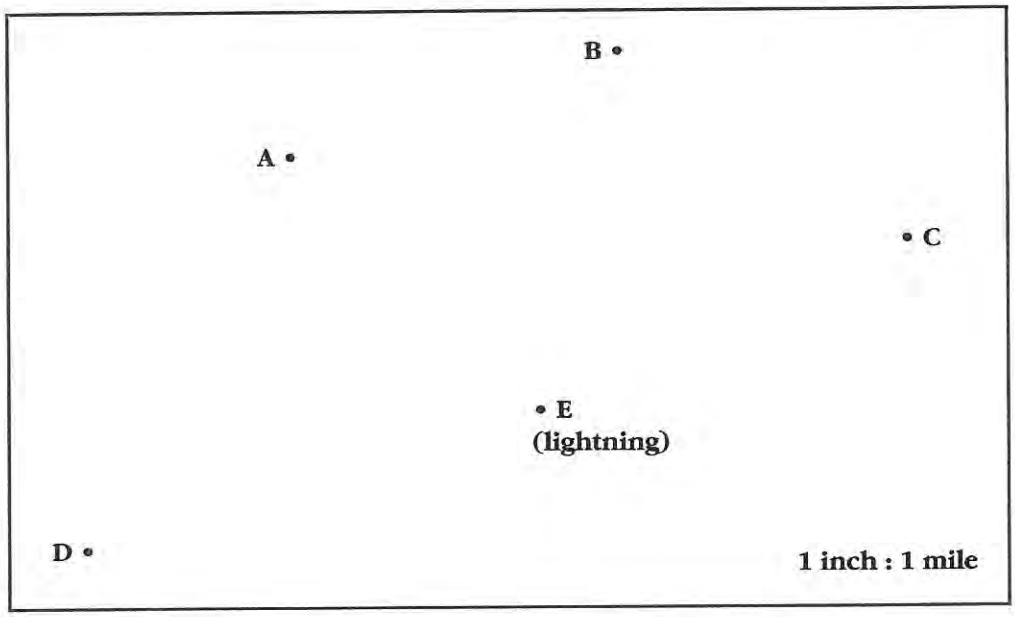
Conditional-Advanced
Grade 12: 92%

SOURCE: I.V.S. Mullis, J.A. Dossey, E.H. Owen, and G.W. Phillips, *NAEP 1992 Mathematics Report Card for the Nation and the States* (Washington, D.C.: National Center for Education Statistics, 1993).

Figure 4.4. ***Sample fourth-grade problem from Measuring Up: Prototypes for Mathematics Assessment published by the Mathematical Science Education Board and the National Research Council***

One way to estimate the distance from you to where lightning strikes is to count the number of seconds until you hear the thunder, and then divide by five. The number you get is the approximate distance in miles.

People are standing at the four points A, B, C and D. They saw lightning strike at point E. Because sound travels more slowly than light, they did not hear the thunder right away.



1. Who heard the thunder first? _____ Why?

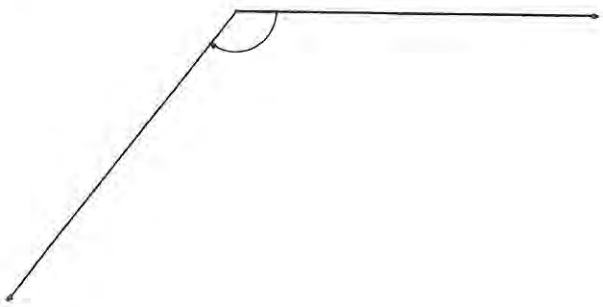
2. Who heard it last? _____ Why?

3. One of the people heard it after 12 seconds.
   Who was it?_____ Explain your answer.

4. After how many seconds did the person at B hear
   the thunder?_____ Show how you know.

Figure 4.5. ***An "exemplar item" used in reporting 1992 NAEP mathematics results that does not exemplify what students at the level can do***

Grade 8 Basic: Example 3



Overall Percent Correct*
Grade 8: 35 (1.9)

Conditional-Basic
Grade 8: 37%

Use your protractor to find the degree measure of the angle shown above.

Answer: 128°

*The standard errors of the estimated percentages appear in parentheses.

Content analyses of the different versions of the descriptions did not reveal significant changes from the St. Louis to the San Diego versions. However, the final version, which was developed by expert consultants for NAGB, was judged to be significantly different from the other two, but more in keeping with the framework. For example, "the distinction among literary, informational, and practical text was incorporated into the descriptors and consistent across levels and grades. The language of the final descriptions more closely paralleled that of the framework, and attempts were made to characterize reading as an interaction among reader, text, and context by including references to grade-appropriate texts."[25] The 12th-grade advanced-level description was cited as an example of the "dramatic" changes that occurred from the first to final version across all grades and achievement levels. (See figure 3.6 in chapter 3.) The list of elements in the St. Louis descriptions included interpretation of figurative language, accessing and revising information across tests and time, and recognizing bias. These elements are omitted in the final version but are replaced by new elements, such as describing abstract themes and ideas; analyzing meaning and form; producing thorough, thoughtful, and extensive responses; producing complex, abstract summaries; applying directions from the text to new situations; and evaluating the usefulness of text. "Only a few ideas such as critically evaluating the text, comparing points of view, analyzing author's stance, and using cultural or historical information provided in text are evident in both versions of the descriptors."[26]

Both content-expert groups saw the final version of the descriptions as a marked improvement over the earlier versions. Teachers who participated in the interview sample also liked the final achievement-level descriptions. They endorsed the high

[25] Op. cit., 193.

[26] Op. cit., 194.

standards of achievement represented and approved of a basic level that was well above minimum competency. However, *the final conclusions from the expert groups in reading were negative. The descriptions had changed too much to be used with the cutscores set in St. Louis, but had not changed enough to be an adequate reflection of the framework (which was treated as synonymous with contemporary professional standards).* In particular, the notion of reading as an interaction among the reader, the text, and the reading situation still was not clearly reflected in the descriptions, and "important factors that might characterize differences in reading achievement such as text difficulty, type of task, or time demands were not included in the descriptors."[27]

An important question with respect to both the reading and the mathematics assessments is the extent to which the NAEP frameworks are similar to what might be developed as national content standards. The cases of the two content areas are somewhat distinct. In mathematics, the 1992 framework as well as the item pool are judged to be somewhat deficient in representing national content standards. In reading, experts see the 1992 framework as being closer to emergent standards. In other respects, however, findings are similar for reading and mathematics. Thus, the reading experts found changes in the descriptions that may have undermined the validity of the cutscores and the match of the assessment content to the descriptions. They also found serious limitations in the item pool. For example, at both fourth and eighth grades, participants in the reading item-mapping study "complained that there were very few items that related to aspects of the advanced descriptors such as the ability to analyze critically or the ability to recognize the use of literary devices."[28] At grade 12, experts found it very difficult to set levels because there were no items to assess aspects of the descriptions such as "critical evaluation" and "integration of historical and cultural information outside the text." The following are direct quotations from experts recorded during the process of matching the descriptions to specific items on the NAEP continuum. They illustrate the kinds of problems that arose when experts could not find items to assess elements of the descriptions:

> (4th grade) "Advanced" says that students should be able to generalize and know how authors compose and use literary devices. You show me where there is an item like that. The only ones that come close are those that refer to personal experiences. And most of those deal with student reference (not author reference).
>
> (8th grade) Not all parts of the descriptors are covered by the test. Students may go as far as to explain literary devices and author's style. But neither the prompt nor the answer key is explicit that they should do that.
>
> (8th grade) Almost no items ask students to critically evaluate what they have read. They might be asked their opinion—but that is not a critique!
>
> (12th grade) I don't understand what is meant by identifying the relationship between author's stance and elements of the text. Show me an item that taps that. I don't think you will find one.

[27] Ibid.

[28] Pearson and DeStefano, Report Three, 406.

(12th grade, proficient) This one part of the descriptor bugs me–the use of cultural and historical information. Are students just supposed to do that (spontaneously)—or should the item call for it? I don't see any that do.[29]

"Should Versus Can" Interpretations of Achievement-Level Results

When NAGB first decided to establish achievement levels for NAEP, they did so after reviewing and rejecting other possible ways to set achievement goals. Specifically, they rejected the idea of setting targets for the proportion of students that should reach certain fixed points on the existing proficiency scale because they saw this way of measuring progress as nonsubstantive. "The fundamental problem with this suggestion is that the proficiency levels are not based on content but on score distributions."[30] By focusing on content, NAGB intended to set standards based on substance and specific subject-matter requirements. However, these substantive claims, which add interpretations to the assessment results, must then be evaluated in addition to the original quantitative properties of the assessment.

Achievement levels are one kind of performance standard. They are intended to establish expectations for what students *should* know and be able to do to attain each level. In this sense the standards are statements of desired rather than actual outcomes. However, once these idealized expectations have been articulated and used to specify cutpoints on the score scale, then it is understood that students who reach the level *can* do what is described at that level. Imagine trying to set standards for climbing a ladder. The standard for being an advanced climber is to reach rung 9 of 10; a proficient climber, rung 8; a basic climber, rung 5. If a climber reaches rung 9, we say that he or she is advanced and can climb to at least rung 9 by definition.

Problems have arisen with initial efforts to report NAEP results by achievement levels because in many instances it is not, or may not be, true that students at a level can do what is described by the level. Note that this is a different problem than the question of whether the levels are too high or too low. Here we are concerned with what students who meet the level can do. Because of ambiguity about the fit of the descriptions and exemplar items to the so-called advanced, proficient, and basic regions of the score scale, the 1992 mathematics report released by NCES said that the achievement levels "attempt to describe what students should be able to do in various ranges on the NAEP scale."[31] This was intended as a needed caveat given the limitations of the current levels. However, in the Panel's view, it is at best ambiguous, if not seriously misleading, and in any case, contrary to the initial intentions of reporting by achievement levels. Furthermore, it is likely to add confusion but not prevent misinterpretation. A preferable decision would have been not to report by achievement levels until the inferences likely to be drawn could be supported by the data.

[29] Ibid., 441.

[30] Roy Truby, "Setting Appropriate Achievement Levels for the National Assessment of Educational Progress" (Washington, D.C.: National Assessment Governing Board, May 10, 1990), 11.

[31] Mullis et al., op. cit., 33.

The inaccuracy of the achievement-level descriptions and exemplars in reporting what students can do at each level occurs because the judgmental process used to establish the levels was not articulated sufficiently with the empirically determined patterns of actual student responses. As seen in the preceding analyses of both the reading and mathematics descriptions, revisions of the descriptions moved them in the direction of professional content standards but away from their respective item pools. Therefore, the descriptions are no longer a very accurate report of the kinds of things that students actually did on the assessment. *The use of achievement levels to communicate standards to the public and to evaluate whether students have attained those standards places new, much more stringent demands on the substantive content and items of the assessment.*

Exemplar items shape public understandings about the nature of subject-matter expectations. If most members of the public are accustomed to viewing mathematics as number facts and algorithms, then seeing a narrow set of exemplar items will do little to convey higher expectations about mathematical thinking and new content standards. As described by subject-matter experts, so-called exemplar items for 1992 were often not exemplary. In many cases, they reflected current practice but not desired standards. Thus limitations in the NAEP item pool, compounded by the necessity to use released items for exemplars, prevented the selection of items that reflect current professional content standards. In the short term, there was no defensible way for NAGB to solve this problem. If more effort had been made to select a substantively better set of exemplars (possibly by releasing more newly developed items), the same problem would have been created that occurred with the descriptions (i.e., moving toward national content standards but away from the assessment items students actually took). The very best NAEP items, chosen to reflect the substance and spirit of the NCTM Standards, are not representative of the NAEP item pool. In chapter 5, reporting alternatives, such as reporting separately for subparts of the assessment selected to reflect current practice versus emerging national standards, are discussed. Ultimately, the dilemma cannot be resolved without substantially redesigning the assessment to support the intended interpretation of results.

In mathematics, the exemplar items were also flawed statistically. As identified by the mathematics experts, several of the reported items did not fit the achievement level to which they were assigned. For example, for one of the basic examples at grade 4, only 49 percent of basic students could do the problem. Thus, it does not illustrate what basic students "can do." At grade 8 only 37 percent of basic students could do one of the basic examples, only 36 percent of proficient students could do a proficient exemplar, and only 42 percent of advanced students could do one of the advanced problems. This kind of misleading example occurred for 5 of the 26 examples used in the 1992 mathematics report. Some inappropriate exemplars had been removed in response to earlier criticism. However, this problem still necessitated an NCES caveat about interpreting the achievement levels as *desired* rather than real attainments at each level.

The problem of misfitting items is relatively easy to remedy by using a statistical criterion (such as requiring 65-percent correct for the category) in addition to substantive ones for selecting exemplar items. Indeed, the Governing Board and NCES significantly remedied one aspect of the "should versus can" problem for the 1992 reading assessment by anchoring the achievement levels with items that students at the various levels could, with high probability, do. While this change is an important

improvement for making accurate interpretations of what students at various levels of achievement can do, the change does not address the problems of the mismatch between the descriptions and cutpoints or the mismatch between elements of the descriptions and the content of the assessment.

Summary of External Comparison Studies

The validity and reasonableness of the achievement levels for interpreting NAEP results depend on the appropriateness of the level cutscores and on the adequacy of the level descriptions. The reasonableness of the cutscores was addressed by a series of external comparison studies; the level descriptions were analyzed by content experts in relation to NAEP frameworks and professional content standards.

Key findings from the external comparison studies are as follows:

1. ***The weight of evidence suggests that the 1992 achievement levels were set unreasonably high.***

Two lines of evidence support this conclusion: (1) field-based evaluations of student performance by teachers and researchers and (2) proportions of students scoring at advanced levels on the SAT and Advanced Placement (AP) examinations. For grades 4 and 8 in both reading and mathematics, the contrasting-groups methodology was used to obtain teacher judgments of student performance. Teachers used the final NAGB descriptions and identified students who fit the descriptions of advanced, proficient, basic, and below basic categories. Teachers identified more students above each cutpoint than were identified by NAEP's achievement levels, suggesting that the NAEP levels were set too high given other evidence of proficiency. Furthermore, individual assessments administered by trained researchers for a subsample of the cases corroborated the teachers' judgments in most cases. Researcher-set cutscores and proportions at or above each level were similar to the teacher results at all levels for both grades 4 and 8 in reading and for grade 4 in mathematics. Only in eighth-grade mathematics were researchers' assessments closer to the achievement-level cutscore classifications than to the teacher classifications.

At grade 12, proportions of students scoring at high levels on the SAT and AP examinations also suggest that the advanced cutpoints may have been set too high in both mathematics and reading. For example, on AP tests in 1992, 2.5 percent and 3.8 percent of high school graduates earned scores of 3, 4, or 5 on the mathematics and English examinations, respectively. (Scores of 3 or better are normally considered adequate to receive college-level credit.) These proportions are quite similar to the proportions of 12th graders reported to be advanced on 1992 NAEP assessments in mathematics and reading. However, given that AP examinations are offered in only 46 percent of U.S. secondary schools, both sets of numbers are arguably underestimates of the proportions of all U.S. 12th graders who are performing at advanced levels.

One external comparison, however, did provide support for the reasonableness of the official achievement levels for eighth-grade mathematics. The percentages of students whose performances were classified as distinguished or proficient on the Kentucky assessment were quite similar to the percentages of Kentucky students scoring at the advanced or proficient levels on NAEP.

2. ***International comparisons do not lead to specific cutscores on the NAEP scale. However, even without achievement levels, international comparisons can be useful for interpreting NAEP results.***

International results suggest that NAEP cutscores for eighth-grade mathematics are not too high in the sense that they are attainable. However, international comparisons do not help to rationalize the choice of specific cutscores because comparisons to different countries and different percentiles would each imply different cutpoints. Should the advanced cutoff be at the 90th percentile of Taiwanese and Korean students? At the 80th percentile? Should it be at the 90th percentile of the top two countries or the top five countries? Each empirical comparison would lead to markedly different cutscores. The international data available for these analyses made a different important point, however. International comparisons are extremely useful for understanding the relative performance of U.S. students. For example, traditional percentile reporting on the IAEP showed that the top 10 percent of 13-year-old Taiwanese students could perform at a level reached by only 1 percent of U.S. 13-year-olds. If percentile equivalences between international tests and NAEP can be established in a defensible manner for benchmarking purposes, it would be feasible to report international comparisons on the NAEP scale. Such comparisons would enhance the usefulness of NAEP results and do not require the use of achievement levels for the results to be meaningful. (This point is further pursued in chapter 5.)

3. ***The item-mapping approach used by content experts revealed deficiencies that appear to be inherent in any item-judgment approach and (consistent with Finding 1) demonstrated that cutpoints are driven higher by lack of advanced items.***

The content-matching study of achievement-level descriptions to items located on the NAEP scale resulted in cutpoints in reading that were higher than the official ones for all but the grade 12 basic and proficient levels. In mathematics, levels identified by experts were systematically more extreme than the official ones, with the expert standards being lower for basic and higher for advanced than the official cutpoints. The quantitative cutpoints derived from these studies were not the focus of the Panel's analyses, however, because the Panel's item-mapping approach may have had as many (and in some cases the same) artifactual influences as the Angoff-set cutpoints. For example, other things being equal, providing experts with items mapped at an 80-percent level probably leads to unreasonably high standards.

Instead, these analyses provided several important insights about the nature of the judgment process. The item-mapping approach led to more consensus-forming discussions; furthermore, it allowed experts to see directly the implications of their deliberations for setting a cutscore. This approach also made it obvious, for both reading and mathematics, that the absence of advanced items forced experts to reach higher and higher on the NAEP scale to try to find items that matched the advanced descriptions.

4. ***Differences among grades in the proportions of students at each level could lead to misinterpretations. In the future, one consideration in standard setting should be defensible grade-to-grade patterns in the proportions of students attaining each achievement level, along with other evidence for grade-to-grade coherence.***

The Panel's examination of grade-to-grade fluctuations in the proportions of students reaching a given level raised issues of possible misinterpretation. For example, achievement-level results cannot be used to infer that middle school mathematics instruction is better than high school instruction simply because there are higher percentages of proficient and advanced students at 8th grade than at 12th grade. Rather than trying to avoid such problems by inserting caveats about what can or cannot be inferred from these patterns, the patterns might better be examined and adjusted in the course of standard setting. A standard-setting methodology that begins with rational consideration of the actual distributions of student performance at different grade levels, and one that incorporates empirical information about score distributions, would obviate the need for any such caveat. NAGB can legitimately eschew "norm-referenced" standards in favor of some kind of "absolute" criterion, but the intelligent use of information about the empirical distribution of achievement does not automatically make a procedure "norm-referenced." To the contrary, this information can help assure the reasonableness and enhance the usefulness of the resulting standards.

Summary Evaluation of the Achievement-Level Descriptions and Exemplar Items

The validity of the final descriptions and exemplar items was evaluated by considering their congruence with content standards, the NAEP frameworks, and the NAEP item pools. Three general problems were identified:

1. ***Current NAEP item pools in reading and mathematics are not adequate for representing emerging national content standards to the public.***

The NAEP item pools, particularly at the advanced levels, are not sufficiently congruent with emerging national content standards for communicating content standards to the public. The quality of released items is especially worrisome. The Panel's content experts were concerned that reporting achievement levels that reify a traditional skills-before-thinking kind of curriculum could "do more harm than good."

2. ***Achievement-level descriptions were changed after the fact and may not be valid for reporting either the cutscores or the assessments.***

For both reading and mathematics, the initial set of achievement-level descriptions that was used to set cutscores subsequently was revised significantly. In the case of

reading, the revisions brought the descriptions more in line with the Reading Framework and professional content standards. In the case of mathematics, the revisions moved the descriptions away from the framework but made them more congruent with the NCTM Standards. In both cases, the revisions were substantial as judged by experts and raise serious questions about the adequacy of the final descriptions to describe the assessment itself or the cutpoints developed using the first version of the descriptions.

3. Some exemplar items are less than exemplary. They do not communicate content standards well and, in the case of mathematics, do not meet statistical criteria.

In addition to the concern that some exemplar items do not adequately match the descriptions (or content standards), several of the items used in the 1992 mathematics report did not have reasonable statistical properties and therefore could not serve as examples of what students performing at a given achievement level can do. This problem has been remedied for the report of reading results by using anchoring statistical criteria to ensure that items selected to typify a given level are indeed items that a majority of students at that level "can do." However, statistical criteria cannot be applied without keeping conceptual criteria in mind. In essence, exemplar items must meet joint criteria to ensure that they reflect both the narrative descriptions and empirical claims. In addition, the items must not be so unusual (compared to other items in the item pool) that they fail to be representative of the assessment as administered. These multiple constraints may not be possible to satisfy with an item pool that was not designed for the purpose of reporting by achievement levels.

NAEP achievement levels were intended to improve the interpretability and usefulness of NAEP results. The 1992 levels have not accomplished this purpose. Instead, flaws in the achievement levels have required burdensome caveats that shroud rather than illuminate the meaning of NAEP achievement data.

5 The Relationship of NAEP to National Education Standards

Findings and Evaluation of the 1992 Achievement Levels

The Panel's evaluation studies were primarily studies of the adequacy of the achievement levels themselves and whether the levels lead to valid inferences about student performance. Our research studies focused on two major questions:

1. Were the processes used to set the levels internally consistent and coherent?

2. Do the final levels appear to be reasonable and valid on the basis of external comparison data and substantive analyses?

The Panel also weighed its findings in light of NAEP's primary purpose, which is to gather scientifically rigorous data to monitor student achievement, and in light of NAEP's relationship to emerging national content and performance standards.

The findings are grouped as responses to the Panel's two primary research questions.

The Process of Level Setting Raises Doubts About Validity and Consistency

- **Different item types led to large internal inconsistencies in judges' ratings.**

Internal data analyses showed that judges were unable to maintain a consistent view of their expectations regarding basic, proficient, and advanced performance. Individual judges did not vary their standards in response to content differences, as might have been expected; instead they appeared to be swayed by seemingly irrelevant features of items such as open-ended versus right-answer items. Individual judges set substantially different cutoffs for items that were empirically the same in difficulty (based on student performance) but different in format. Within-judge differences for the same achievement level, say, fourth-grade basic, differed by as much as the average cutscores for adjacent grades—say, fourth- versus eighth-grade basic.

- **The Angoff procedure is fundamentally flawed because it depends upon cognitive judgments that are virtually impossible to make.**

The Panel had access to internal consistency data of a sort that have not been available previously in the standard-setting literature. The Panel found that participants were willing to comply with the demands of the Angoff procedure without having a

workable understanding of how conceptions of borderline students at each of the levels should be translated into item p-values. In fact, the Angoff judgment task requires panelists to make difficult hypothetical judgments that are virtually impossible to make. The Panel concluded, therefore, that the approach is fundamentally flawed for setting achievement levels on NAEP. Furthermore, focusing on ratings of individual items using the Angoff or any other item-judgment method does not allow judges to develop integrated conceptions of performance standards.

- **There was no evidence that judges reached agreement or developed a consensus.**

In both reading and mathematics, the mean recommended cutpoints stayed roughly the same from Round 1 to Round 3 of the Angoff procedure, and, more important, there was relatively little decrease in the variability among judges. Furthermore, the variability in judges' desired cutpoints, even at Round 3, was large. At the end of the process, the standard deviation of judges' ratings was typically 14 points. This means that even the middle two-thirds of the judges disagreed—about where the cutpoint should be for a level—by as much as 28 points, which is approaching the within-grade standard deviation of student scores of 40 points. From observations of the reading level-setting process, it was found that discussion of the kind needed to form a consensus occurred in only one of the fourth-grade groups. Thus, it cannot be argued that representative groups developed consensus standards that should be adhered to, regardless of technical flaws in the procedures.

- **Initial failure to tie achievement levels to NAEP frameworks undercuts the validity of the cutpoints.**

In reading, participants who set the levels were largely unfamiliar with the NAEP framework. Therefore, the initial achievement-level descriptions were not adequately connected to the framework but relied more on judges' personal experience. Lack of fit to the framework was remedied subsequently by making significant changes in the descriptions. Changes in the descriptions were also made after the fact in mathematics, making the descriptions more congruent with current thinking about national content standards in mathematics but less congruent with the assessment. These changes in the descriptions necessarily raise serious questions about the validity of cutpoints set with one set of descriptions and interpreted with another.

NAGB's contractor made an effort to establish the correspondence between the cutpoints and the final descriptions by surveying participants to see if they thought their ratings would have been changed by the change in descriptions. Unfortunately, the question cannot be answered well by this type of post-hoc speculation. The Panel's experts in both reading and mathematics had access to both sets of descriptions and concluded that they were substantially different, reflected different conceptions of subject matter, and probably would have lead to different item ratings. For example, at the advanced levels in both subject areas, experts had to move cutpoints higher in response to the lack of items matching assertions made in the final descriptions.

External Comparison Studies and Content Analyses Raise Serious Questions About the Validity of the Cutscores and Achievement-Level Descriptions

There are no absolute external criteria for judging the validity of the NAEP achievement levels. However, various external comparisons can be used to evaluate the reasonableness of the levels. The kind of evidence needed to evaluate validity is like that used to establish test validity for any difficult-to-measure construct. First, are the patterns of relationships with other measures like those that would be expected based on theoretical understandings? Specifically, do other sources of data lead to similar conclusions about the proportions of students in each achievement category? A second question pertains to the substance of the achievement levels as reflected in the descriptions and exemplar items. Do they communicate accurately what students should be able to do to be considered advanced, proficient, and basic? For our analyses, the Panel used the results of several external validity studies and data comparisons along with analyses of the levels by subject-matter experts. (See chapter 4.) Study results can be summarized briefly as follows:

- **The weight of the evidence suggests that the 1992 achievement levels were set unreasonably high.**

Contrasting-groups studies, endorsed by NAGB staff as appropriate empirical verification procedures, were conducted in both reading and mathematics at grades 4 and 8. Using the NAEP descriptions, teachers classified students who had also taken the NAEP assessment into advanced, proficient, basic, and below basic categories. For a subsample of students, teacher judgments were followed up by individual assessments administered by trained researchers. These classifications of students were translated statistically into cutpoints on the NAEP scale. In both reading and mathematics, teachers consistently identified more students as advanced, proficient, and basic than were found to be above the NAEP cutpoints. In reading, teacher classifications were consistently corroborated by researcher evaluations. In mathematics, this same result was found at grade 4 but not at grade 8 where the researchers' ratings for the advanced and basic categories agreed more with the NAEP classifications.

Because educators and policymakers have some experience with the meaning of high scores on the SAT and Advanced Placement examinations, percentages of students earning above 550 on the SAT Verbal test, above 600 on the SAT Mathematics test, or a "3" or better on AP tests (expressed as a percentage of all graduating seniors) provide a useful basis for estimating what proportion of 12th-grade students should be considered advanced. For example, 7.5 percent of high school graduates scored above 600 on the SAT Mathematics test, and 5.8 percent scored above the corresponding level on the SAT Verbal test. These percentages should be treated as underestimates because many able students take the ACT rather than the SAT. On Advanced Placement examinations, 2.5 percent of high school graduates qualified for college credit (scoring 3 or better) in mathematics, and 3.8 percent qualified for college credit in English. Again, these are underestimates of the total percentage who might score this high given that AP examinations are offered in only 46 percent of U.S. high schools. Taken together, the SAT and AP data suggest that higher percentages of 12th graders should have been classified as advanced on NAEP.

- **International comparisons do not lead to specific cutscores on the NAEP scale; however, even without achievement levels, international comparisons can be useful for interpreting NAEP results.**

International data available on 13-year-olds in mathematics led to ambiguous results because equating to NAEP by two different statistical procedures led to different estimates of percentages of students at each of the achievement levels. International data are useful for evaluating the performance of U.S. students even without equating to achievement levels. Traditional reporting of international assessment data showed that the top 10 percent of Taiwanese students could perform at a level reached by only 1 percent of U.S. 13-year-olds in 1991. If percentile equivalences between international assessments and NAEP could be established in a defensible manner, then it would be possible to report useful international comparisons on the NAEP scale.

- **The item-mapping approach used by content experts revealed deficiencies that appear to be inherent in any item-judgment approach and demonstrated that cutpoints are driven higher by lack of advanced items.**

Panels of subject-matter experts were convened in reading and mathematics and asked to evaluate (1) the appropriateness of the level cutpoints, and (2) the quality of the descriptions. To facilitate their work, expert teams were provided with "item maps," with items located along the NAEP scale at the point where 80 percent of students were expected to get the item right (after correcting for guessing). In mathematics, the plausible ranges for cutpoints identified by experts were consistently lower than the official basic-level cutpoints and higher than the official advanced-level cutpoints. In reading, with two exceptions, the cutpoints set by experts were higher than the official cutpoints.

Taken together, findings from the Panel's several studies provided insights about how judgmental procedures may affect the determination of cutpoints. These kinds of issues or patterns have not been identified or studied previously in the research literature. Studies that looked at integrated evidence of student work (such as whole-test booklets and classroom work) tended to identify more students' performance as advanced and proficient than item-by-item judgment procedures. The Panel speculated that this was likely to occur because judges accepted different combinations of item performance (not all at the same level) as evidence that students had attained levels of achievement consistent with the descriptions. In contrast, the Panel observed that when judging exemplar papers or items on an item-mapped scale, judges appear to assume perfect reliability of performance across tasks of similar difficulty. This assumption has the effect of driving up cutscores. Most important, in the absence of items representing conceptually challenging content, experts attempting to set cutscores faithful to the achievement-level descriptions set higher and higher standards on low-level content.

- **Current NAEP item pools in reading and mathematics are not adequate for representing emerging national content standards to the public.**

Post-hoc revisions attempting to make the achievement levels more consistent with professional content standards resulted in achievement levels that were inconsistent with the assessments and the cutscores.

A serious concern raised by subject-matter experts about the validity of the levels pertained to their substance, not the quantitative cutpoints. In mathematics, the concern arises because both the 1992 framework and item pool fail to adequately represent the NCTM Standards.[1] Consequently, neither the narrative descriptions nor exemplar items are adequate to use in communicating national content standards to the public. In reading, the NAEP framework appears to be further along in representing current professional and research-based conceptions of the subject, but the items in the assessment were not developed to be a complete representation of the framework. The final narrative descriptions in reading do a reasonably good job of conveying professional expectations but in some cases are not accurately linked to the item pool. Thus the verbal descriptions of what students can do at each level do not correspond to what they actually did on the test.

The Panel's Recommendations

As noted early in this report, members of the Panel endorse the need to establish challenging education performance standards that exemplify achievement expectations. However, as summarized above, the Panel found many problems with the manner in which the Governing Board set achievement levels for the 1992 mathematics and reading assessments. The standards set must be defensible to assure that education policy decisions based on them are sound. The Panel concludes that the process by which the 1992 achievement levels were set cannot be defended given the results of the studies reported above.

The Panel is aware that state assessment directors are in favor of reporting NAEP results using achievement levels. The Panel understands why policymakers might also favor standards-based reporting. Standards-based reporting has appeal because it communicates not only what the nation's students know, but what they *should* know as well. Reported results can be either a source for pride or a motivating force to further education reform efforts. But the Panel also believes that setting defensible education standards is a long term judgmental and empirical process: such standards cannot be set in a few weeks, or even in a few months. For these reasons, the Panel has divided its recommendations into two sets—short term and long term.

[1] A new mathematics framework that is better aligned with the NCTM Standards has already been developed for use in future assessments.

1. **Discontinue Use of the Angoff Method.** The Panel recommends that use of the Angoff method or any other item-judgment method to set achievement levels be discontinued. As the Panel's studies demonstrate, the Angoff method approach and other item-judgment methods are fundamentally flawed. Minor improvements, such as allowing more discussion time or providing instructions about guessing, cannot overcome the nearly impossible cognitive task of estimating the probability that a hypothetical student at the boundary of a given achievement level will get a particular item correct. Furthermore, the Angoff method does not allow for an integrated conception of subject-matter proficiency.[2]

2. **Discontinue Reporting by Achievement Levels as Used in 1992.** The Panel urges NCES and NAGB not to report the 1992 NAEP results by achievement levels. The descriptions and exemplar tasks are not adequate for reporting to the public what students should be able to do. Furthermore, the assessment content does not measure up to that expected for the emerging national content standards. Cutpoints set in both 1990 and 1992 warn us that attempts to establish performance standards with inadequate content run the risk of setting high levels of performance on low-level content. NAEP content must be expanded substantially to reflect emerging national content standards. When the content changes, trends based on the old levels will be potentially misleading. Thus it only makes sense to wait until national content standards are available and then to follow a more coherent process for developing performance standards in conjunction with content standards.

   This recommendation may appear to be extreme. It is contrary to the desires of Governing Board members, who represent a broad array of lay and professional perspectives, and it is contrary to the wishes of state-level users of NAEP data who would like to have agreed-upon national performance standards. Objections such as the following will certainly be made: Why not use the current achievement levels for reporting until national content standards are available? If there is no perfect standard-setting method, and no absolute validity criterion, aren't the current achievement levels good enough? Furthermore, it may take

[2] Angoff procedures may or may not be defensible in other contexts (e.g., setting minimum standards based on all-or-none judgments about essential knowledge for a specific vocation). Based on the Panel's findings concerning internal and external validity, the NAE Panel is understandably skeptical. However, the Panel would not want the discussion here to be taken in any way as an indictment of other standards established by such methods.

several years before results could be reported using the more iterative and comprehensive approach suggested by the Panel later in this chapter.

In response to these objections, it should be noted that the Panel's recommendation is based on the indefensibility of the internal-process data and on findings from the subject-matter studies. If exemplar items and descriptions are not exemplary representations of current thinking in the fields of mathematics and reading, then reporting on this basis could jeopardize the larger effort to create national content standards. Examples that use low-level skills could set educational sights too low and create the wrong impression of the goals for which the nation is striving. In addition, the spotlight that is now focused on the achievement levels in subject-matter communities could harm the credibility of NAEP.

The Governing Board has worked to get the message out quickly that the achievement of America's students is seriously wanting. This was a key reason for setting short timelines for reporting. The Governing Board also believed that the best way to make this message understood was through the use of achievement levels. They believed that achievement levels would make results more interpretable. Yet because of technical difficulties in setting the achievement levels, their use has substantially *slowed* the reporting process. The 1992 mathematics results were not released until April 1993; the 1992 results for reading will not be out until September 1993; and the 1992 writing results will not be released until late 1993 or early 1994, nearly 2 years after the assessment was completed.

The confusion that occurred with the 1992 mathematics reports as to whether students who reached a given level could in fact do the work described as defining that level (the so-called "should versus can" issue) strongly suggests that interpretation was harmed by the use of the achievement levels. The Governing Board and NCES have significantly remedied one aspect of this particular problem for the 1992 reading assessment by combining the scale-anchoring technology with the use of achievement levels. While this change improves the ability to interpret the meaning of the levels, it does nothing to address the problems of the internal consistency of the process, or the problem of the mismatch between the descriptions and the cutpoints.

The Panel is aware that plans for the NAEP assessments in 1994 are well underway. The Governing Board has determined the subject areas to be U.S. history, geography, and reading. Current plans are to assess in all three subjects at the national level, with a state-by-state assessment in reading only.[3]

[3] Appropriations have not yet been established for FY-94, and therefore the scope of the assessment with respect to which grade levels to assess is not yet fixed.

The Governing Board has issued a request for proposals for the development of achievement levels for the U.S. history and geography assessments. The Panel recommends reconsideration of the intent to develop achievement levels, particularly with the use of the Angoff or any other item-judgment method for the 1994 assessment. The Panel also recommends against the use of the 1992 achievement levels in mathematics and reading as baselines against which to make comparisons in future assessments, given the flaws found with the Angoff process for the setting of achievement levels.

3. **Invite Content Experts, Business Leaders, and Standards Committees to Comment on the Meaning of NAEP Results and Desired Performance Standards.** The Panel recommends that beginning with the 1992 reading assessment, NAGB contribute to the dialogue about national content and performance standards by inviting different groups, such as subject-matter experts and business leaders, to study NAEP results closely and produce an evaluative commentary. Such groups could even be encouraged to consider the question "How good is good enough?" By adding policy interpretations after the reports are released to the public, rather than building them into the score reporting, it becomes possible to have a more informed debate about competing perspectives. It might even be possible to focus public discourse on substantive issues about what should be taught, rather than emphasizing a quantitative horse race. (For example, the public should know what is covered in eighth-grade mathematics curricula worldwide, rather than simply knowing that 13-year-old U.S. students ranked 14th among those countries that participated in the International Assessment of Educational Progress [IAEP].) Policy commentaries accompanying NAEP but separate from the data could also suggest policy intervention at the state and local levels needed to effect changes in student performance.

   In particular, the Panel believes that the development of national content standards and the development of NAEP frameworks and performance standards would each benefit from a formal dialogue between NAEP content committees and the groups charged with developing national content standards. Although the development of content standards should precede the development of performance standards, content standards cannot be developed purely in the abstract without concrete examples of instructional activities and student work. Therefore, the Panel specifically recommends that the Governing Board invite national standards committees in each subject area, such as the National Academy of Sciences/National Research Council science content committee, to examine NAEP results closely, at the level of specific tasks and items, and report on their meaning. Exemplar items could be used to illustrate desired performance at various levels. The Panel expects the benefits of such an exchange to accrue on both sides. NAGB would gain an evaluative interpretation useful in the description of NAEP results; the

standards group would gain insight into the measurement of performance standards. Finally, content standards committees would also benefit from access to broadly representative examples of student performance and (indirectly) evidence of current teaching practices.

4. **Publish Achievement Levels in 1994 Separately from the Official NAEP Reports, and Report These as Draft or Developmental.** The primary means of reporting in 1994 should be by average NAEP scale scores and anchor points. However, the Panel also favors the development of thoughtful standards-based reporting. As is pointed out in the section on long term recommendations, the establishment of performance standards based on the emerging content standards cannot be done in the short term. Nevertheless, the Panel recommends that efforts to establish experimental or trial achievement levels in reading, history, and geography begin in 1994. New standard-setting procedures should be developed and implemented on a trial basis. Before attempts are made to set cutscores for the 1994 assessments, descriptions, consistent with the NAEP frameworks, should be agreed upon. Rather than rely on a single procedure, potential achievement-level cutscores should be developed using several approaches, with followup efforts then made to reconcile differences and arrive at defensible standards. Based on insights from its evaluation studies, the Panel recommends that at a minimum *all three* of the following approaches be used: (1) contrasting-groups field-based studies, (2) an item-mapping procedure, and (3) a total-student-performance (whole-booklet) mapping procedure. The item-mapping approach was used by content experts in the studies described in chapter 4; however, in the future a more appropriate mapping criterion should be used, such as the location where 65 percent of students get the item right. A corresponding procedure could be used to map sample pupil booklets to the NAEP score scale and would enable judges to examine more integrated evidence of student performance. As part of the process, exemplar items should be identified that meet both conceptual and statistical criteria.

Efforts to set achievement levels using these new strategies must be carefully evaluated, and it is essential that the results be viewed only as a research and development effort. Therefore, the achievement-level results should be released by NCES as part of its ongoing Research and Development series of publications, preferably sometime after the release of the official NAEP results. To avoid previous problems that made it necessary to revise the achievement levels after the fact and recompute trend data, achievement levels for each subject area should be reported in draft form for at least one assessment cycle before being adopted as the official levels. Such a process, even in the short term, would allow adequate time to respond to criticism and evaluation findings.

The Panel believes this recommendation is consistent with the reporting mission of a statistical agency. When a statistical agency considers adding a new index or changing an old one (e.g., the current move by the Bureau of Labor Statistics to use a revised measure of unemployment), it is customary to implement the change on a pilot basis and to collect data evaluating the change prior to institutionalizing it. NAEP routinely tries out and evaluates changes before incorporating them. Examples include the use of estimation items in mathematics and the special study in oral reading in 1992. *Reports of draft achievement levels should include data from empirical studies to evaluate the levels for internal coherence and for reasonableness in comparison to external evidence of students' performance.* The Panel also suggests the use of commentary from stakeholder groups from history, geography, and reading in the case of the 1994 developmental effort.

Achievement levels should not become the primary method of reporting NAEP results until they have been through the substantive and empirical evaluations suggested in the long term recommendations at the end of this chapter. However, examples of other types of benchmarks that have interpretive value and might be developed in the short run are suggested in sections below.

5. **Use 1990 and 1992 Percentile Scores to Monitor Achievement in Future Assessments.** Reporting NAEP results using performance standards closely aligned with national content standards is desirable, but the content standards are not yet established; therefore, NAEP results cannot be reported using them. In the meantime, ways of reporting NAEP are needed to help the public and policymakers (a) see trends in performance and (b) understand changes in various levels of performance. Therefore, the Panel recommends that baseline performance could be set at three levels using thoughtfully chosen percentile scores. NAGB, in consultation with NCES, should decide the three percentile cutscores. The Panel suggests the 95th, 75th, and 25th percentiles for a base year could be used as benchmarks against which to measure future progress. This procedure could be implemented on the 1990 NAEP mathematics assessment. Doing so would allow the Governing Board to trace progress in mathematics from 1990 to 1992 and then again from 1992 to 1995, or whenever the next mathematics assessment is administered. The Governing Board would decide what percentage of students should reach the three target levels in future assessments, and NCES would report progress against meeting those targets. For example, in setting a target for the year 2000, the Governing Board might decide that 20 percent of the students should be above the 1990 95th percentile benchmark in mathematics. As an alternative or addition, comparisons using international data (see recommendation 6 below) could be used.

Percentile scores have the advantage of being simple, technically defensible, and easy to explain to the public. To further enhance interpretation, NCES might decide to produce anchor-point descriptions to accompany the cutpoints. Based on the current achievement levels and results from the external comparison studies, the 75th and 95th percentiles are clearly in the regions of the score scale that represent proficient and advanced performance, respectively. To ensure that performance at the 75th percentile represents "competency in challenging subject matter" and is therefore appropriate for measuring progress toward Goal 3 of the national education goals, NCES could examine anchor descriptions and anchor items *prior* to selecting the percentiles to be used. The substantively most defensible percentile to represent "challenging subject matter" could be used to choose the 70th, 75th, or 80th percentiles for reporting; however, *the same percentile should be selected for all three grades to avoid the potential for misinterpretation of grade-to-grade trends found with the achievement levels.*

The Panel strongly suggests setting one of the baseline percentile scores at the low end of the distribution (e.g., 25th percentile), because it believes it is important to mark the progress of the large number of students that are currently performing poorly on NAEP. The use of the three baseline cutpoints at roughly the 95th, 75th, and 25th percentiles would allow the nation to monitor the performance of its students in three meaningful regions along the NAEP scale.

Baseline cutscores can also be set for the 1992 reading assessment, allowing NAGB to set targets for the 1994 reading assessment. Similarly, baseline performance in history and geography, as well as target gains for future assessments in these content areas, can be determined after the 1994 assessment is administered and scored.

6. **Use International Comparisons to Set Benchmarks for U.S. Performance.** As shown in chapter 4, comparisons between the mathematics achievement of U.S. eighth graders and the achievement of students from other nations can also provide useful information. Therefore, in addition to using percentile scores on NAEP, the Panel recommends that NAGB and NCES compare the performance of U.S. students with that of students in other nations. While the details for making these international comparisons should be determined by NCES, one suggestion is to use international percentile results as benchmarks if defensible percentile equivalences can be established between NAEP and international assessments. For example, the score corresponding to the 90th percentile for the combined results of, say, the four highest-scoring countries (e.g., on the 1991 IAEP or the upcoming Third International Mathematics and Science Study) could be identified as advanced performance, and the corresponding percentile rank for U.S. students could then be determined. The NAEP score corresponding to the latter percentile rank could then

be used as a benchmark to track the progress of U.S. students on future NAEP assessments. That is, the percentage of students scoring at the benchmark level or above becomes the baseline against which to make future comparisons. It would be within the purview of the Governing Board to set target percentages to be reached in future assessments as goals. Stated differently, NAGB could target the percentage of students that it believes should be "world class" in mathematics and science by the year 2000.

The Panel recognizes the limitations of international comparisons. Countries differ with respect to the coverage of students sampled, and the IAEP mathematics assessment itself has limitations (e.g., the item pool is limited in the number of items that tap higher-order thinking). Furthermore, at this point, the only subjects for which we will have recent international comparisons are mathematics and science. In spite of these limitations, a benchmark such as that described above *does* provide a meaningful way to track progress for Education Goal 4—"By the year 2000, U.S. students will be first in the world in science and mathematics achievement."

7. **Work with the National Education Goals Panel to Develop a Way to Use NAEP Results to Measure Progress over the Decade of the 1990s.** The current achievement levels are flawed and therefore cannot be used to measure progress towards the nation's education goals. The Panel recommends that NAGB and NCES join with the National Education Goals Panel to work out a mutually acceptable way for reporting the results over the decade of the 1990s (i.e., until such time as performance standards based on certified national content standards can be developed). The Panel has suggested several possible strategies for reporting NAEP results in the short run. NAGB, NCES, and the Goals Panel should agree on a common method for reporting progress. A common method is important because, as the Panel noted in its evaluation of the 1990 Trial State NAEP, it was confusing for the public to have the same results reported in different ways by NCES and the Goals Panel.

8. **Implement Within-Grade Score Reporting.** At present, performance for students in grades 4, 8, and 12 is reported on a single 0-to-500-point scale. In addition, the score scale is constructed (in standard deviation units) so that, by definition, even 4th graders rarely score below 150 and 12th graders rarely score above 350. Therefore, the scale is effectively compressed into a range of only 200 points (from 150 to 350). As part of its argument for achievement levels, NAGB expressed concern that use of a single scale obscured variation within grade and made it more difficult to make evaluative judgments. "For example, with

only one common scale for mathematics, almost no 4th grader will ever be at the advanced level even though a sizeable percentage of 4th-grade students may be doing what is advanced work for the 4th grade."[4]

The Panel considers the use of a single vertical scale to be problematic on other grounds. In its previous report, the Panel noted that reporting a single score in mathematics was technically misleading and would not serve the purpose of improving mathematics instruction in areas such as geometry, algebra, measurement, and statistics. The decision to have a subject-matter scale span multiple grades complicates the meaning of a composite score because weighting of the composite changes with grade level. Even within content strands, the assumptions of the scaling methodology are strained by combining performance across eight grade levels. Does a score of 250 have the same meaning for a fourth and eighth grader, for example?

The NAGB achievement levels address one aspect of this question by adjusting standards to grade-level expectations. In mathematics, the 1992 achievement levels suggest that a fourth grader with a score of 250 is proficient (just barely) and an eighth grader with the same score would be (just barely) below basic. But would their performance be similar as suggested by earning identical scores? Would the implications for instruction be the same? Or might there be qualitative differences between a fourth grader's and an eighth grader's understanding of mathematics? The Panel does not have definitive answers to these questions. However, a comparison of mathematics items mapped to the NAEP scale suggests that equivalent scores do not have the same meaning across grades.[5,6] In any case, the answer could well change as NAEP exercise pools become more comprehensive. During the course of the evaluation studies, the Panel heard many comments from subject-matter experts proclaiming that a strength of the achievement levels was their focus on describing performance within grade level. In the absence of achievement levels, exemplar items based on selected within-grade anchor points could be used to describe what performance at various levels looks like. The Panel also believes that within-grade scales will make it easier for subject-matter experts to comment meaningfully on the quality of NAEP performances. For these reasons, the Panel recommends moving to within-grade scaling effective with the 1994 assessment.

[4] Roy Truby, "Setting Appropriate Achievement Levels for the National Assessment of Educational Progress" (Washington, D.C.: National Assessment Governing Board, May 10, 1990), 8.

[5] G.W. Phillips, I.V.S. Mullis, M.L. Bourque, P.L. Williams, R.K. Hambleton, E.H. Owen, and P.E. Barton, *Interpreting NAEP Scales* (Washington, D.C.: National Center for Education Statistics, 1993).

[6] The Panel is primarily concerned with the equivalence of scores across grades for the composite scale. We are aware that the appropriateness of the vertical scale is checked within strand by verifying that cross-grade items scale for students in different grades.

The eight recommendations detailed in the section above are short term in the sense that the Panel believes that they can and should be implemented in the near future. Importantly, the Panel believes that these recommendations should remain in place into the long term as well, with the possible exception of recommendation 5 regarding percentile scores which may be superseded by performance standards of the type described below.

Long Term Recommendations for Developing National Performance Standards Congruent with National Content Standards

As national content standards are developed and certified, the Panel believes it imperative that performance standards on NAEP be linked to them. This is a time-consuming process. The Panel also believes that the development of such performance standards requires a knowledge base for understanding the meaning of various levels of performance. A knowledge base of this sort cannot be developed quickly enough to be available for the next assessment cycle. For these reasons, the Panel believes that the Governing Board must also take a long view as it seeks to establish performance standards. With this perspective in mind, the Panel now presents long term recommendations.

1. **Develop Content Standards and Performance Standards in an Iterative Process.** *Content standards* are analogous to broad curricular frameworks. They identify the subject-matter knowledge and applications that are the goals of instruction. *Performance standards* specify what students must be able to do to reach different levels of competence within those content domains. In both current and traditional views of assessment, the normal process of test construction is to first specify the content domain and then develop representative tasks and performances. Because performance standards must include descriptions of what students at each level should be able to do, along with a quantitative score defining that level, the Panel recommends that performance standards be developed in close coordination with emerging national content standards and assessment tasks. As these content standards become available in each discipline, performance standards can then be created consistent with those content goals. Answering the question of how much students should know without first establishing content standards is virtually impossible. Therefore, measuring progress toward national goals in all NAEP subject areas must await the development of additional national content standards.

   Although the logical order of development is from general content frameworks to assessment tasks to performance criteria, the Panel does not wish to convey that this is a strictly linear and sequential process. In fact, thinking about content standards can be informed by studying sample assessment tasks and student

responses. Decisions about where to set performance standards might well prompt development of additional assessment tasks to clarify distinctions in students' achievements. An ideal process would not set each stage in stone before consideration of the next and would allow time for iteration and revision in response to both substantive critiques and empirical data.

2. **Establish a Standing Subject-Matter Panel for Each Subject Area.** In its last report, the Panel argued that fragmentation of assessment development could be alleviated if subject-matter panels were established with ongoing responsibility for overseeing all phases of the assessment. This same idea is even more important in the context of developing performance standards. In keeping with a consensus approach, the Panel recommends the establishment of standing subject-matter panels that are broadly representative of classroom teachers, curriculum experts, members of the lay public with expertise in the subject area, and researchers. The standing panel in each subject should participate in the development of new frameworks. As part of their deliberations, the standing subject-matter panel should consider explicitly how to incorporate national standards, how to reflect current practice, and whether these perspectives should or should not be combined in a composite score. The same panel would review item development, agree on narrative descriptions of performance standards, and work on selecting representative tasks until there was both logical and statistical correspondence between content standards, performance standards, and illustrative tasks.

   Finally, standing subject-matter panels should become increasingly knowledgeable about the issues of standard setting. The panels should be responsible for making revisions in draft frameworks and standards in response to various knowledgeable constituencies. In addition, the panels would be the appropriate bodies to weigh different sources of information to arrive at performance standards. For example, if predictions about college entrance suggested one level for standards, and international comparisons suggested another, given that no standard-setting method produces a "true" standard, it is reasonable to weigh both substantive considerations and empirical evidence before making a final decision. Ad-hoc consensus groups that have not had sustained experience with the assessment frameworks and content are not in a good position to make these kinds of judgments; standing subject-matter panels would be.

3. **Address Important Conceptual Issues.** The Panel recommends that subject-matter panels working with teams of developmental specialists be asked to address, over a period of time, conceptual issues that have thus far not been considered as part of the effort to specify achievement levels. These issues have to do with underlying assumptions regarding the nature of proficient and advanced performance in each subject area and the development

of proficiency across grade levels. To date, there has been no theoretical or developmental model underlying NAEP scales. Rather, the scales have been created statistically, without evaluating the implied developmental model that underlies them.

Creation of performance standards to accompany national content standards will require reckoning with the following questions. Does advanced performance imply going further into the next grade's curriculum, or can it mean deeper and richer understanding of typical grade-level topics? If accelerating mastery becomes the model of excellence, would highly visible national standards foster superficial treatment of content? What developmental model best characterizes grade-to-grade progressions? Does empirical evidence support the presumed model? It would be important to know, for example, whether students can make the implied grade 4 to grade 8 gains under conditions of good instruction. Are there different models for each achievement level? Do advanced fourth graders become advanced eighth graders, and so forth? Is it possible for a basic fourth grader to become an advanced eighth grader? If so, under what conditions? [7]

If the purpose for having standards and highly visible assessment results is to influence policy and practice, then it is essential that forethought be given to the message that assessment results will convey. And the message itself depends upon having a better understanding of the nature of types of performance.

4. **Empirically Evaluate Achievement Levels Before Making Them Operational.** After subject-matter panels have developed performance standards for NAEP, the Panel recommends that the standards be rigorously evaluated before making them an operational part of the reporting of NAEP results. In particular, the Panel recommends that the levels be validated using the contrasting-groups method that was used to evaluate the current set of achievement levels. Performances should be examined to ensure that the observed patterns confirm predictions based on the explicit or implicit conceptual models used to formulate the performance standards.

   Under current procedures for setting achievement levels, panels of experts or judges define an achievement level by listing the kinds of things students at that level should be able to do. There is no direct, empirical check that the things listed are equally difficult, or that taken together they represent a coherent point in any typical student's educational development. In mathematics, for example, many students might reach the advanced level in some subareas long before others. This type of validity problem must be addressed empirically as well as conceptually. For these reasons it is essential that standards be rigorously evaluated.

[7] Note that these questions are still relevant even with within-grade score reporting.

5. **Recognize the Need for a Multiyear Process for the Development of Performance Standards.** Future efforts to develop national consensus standards should not rely on highly constrained meetings and timetables. Instead, a national consensus process not unlike the 3-year effort to develop the NCTM Standards should be established.

   The NCTM Standards were developed iteratively. A committee of classroom teachers, supervisors, teacher educators, mathematicians, and researchers met in working groups and drafted content standards. The draft was widely circulated in the field; a copy was mailed to each NCTM member. Written comments were received, and hearings were held at regional and national meetings. Respondents were asked to comment on the vision of school mathematics represented in the standards, the quality of each standard, and consistency across standards; and they were invited to submit examples and vignettes.

   A process for developing consensus requires that participants from various constituencies have an opportunity to see what proposed standards will look like. This means that the product available for critique must be the first draft of an integrated whole and have all the essential pieces: content standards, narrative descriptions, sample tasks, and test questions, as well as proposed performance standards. This should be followed by a genuine opportunity to change the proposed standards in response to feedback. If critiques or empirical evidence suggest the need for revisions, timelines must allow opportunities to make them.

6. **Provide for a Stable Basis for Comparison as Well as for Evolutionary Change.** Just as current curriculum frameworks are out of date, national content standards, such as those being developed under the aegis of the National Academy of Sciences/National Research Council, the National Council of Teachers of English, and other groups, will need to be revised from time to time. Changes in assessment content and performance standards must necessarily follow. However, for national content standards to be feasible and useful, they must not change every 2 to 3 years. The Panel recommends a cycle of implementation, feedback, and revision that takes place over, perhaps, an 8- to 10-year period. Stability over some sustained period of time is important substantively, technically, and educationally. An 8- to 10-year cycle is long enough to establish meaningful trend lines, but short enough to accommodate the natural evolution that occurs in subject-matter fields.

   Standards take time to affect practice. The NCTM Standards have been extolled for 3 years and still have not reached most classrooms. Teachers' conceptions and instructional repertoires do not change without retraining and support for curriculum development. For example, teachers and students in elementary schools are accustomed to spending most of their time in

mathematics on computational problems. When the standard of having children communicate mathematically is introduced, having to explain an answer will be initially foreign and difficult. Systemic change requires changes at all levels, in teacher training institutions, in shared understandings between adjacent-grade teachers, in agreements between high schools and colleges, in the materials supplied by publishers and professional associations, and in parental attitudes. For standards to create shared understandings, there has to be time for concrete experiences with them and for discussions about their influence on learning.

Stability of standards is also essential to the integrity of assessment results. Cutpoints like the NAEP achievement levels are known to be arbitrary and imperfect. Even if performance standards were supported by a technically impeccable process and well-reviewed substantive arguments, they would still represent a choice, not an absolute distinction on a performance continuum. Given the arbitrary nature of performance standards, meaningful comparisons across time can only be made if the standards are held constant during that period of time. If standards were changed directly by a new judgmental process or indirectly by substantial changes in test content, then it would not be valid to interpret differences in the percent of students at each level from year to year. Again, an 8- to 10-year cycle for changing standards seems reasonable.

Implications for the Design of NAEP

The recommendations that the Panel has presented have important implications for NAEP itself, implications about which much serious thought and discussion must occur. These implications are neither simple nor easily resolved, but failing to consider them could have a negative impact on the value of NAEP as the Nation's Report Card. In the section that follows, the Panel identifies four sets of issues that the Governing Board and NCES should explore and consider.

1. The Setting of Achievement Levels Needs to Be Integrated with Other Efforts to Develop National Education Standards.

The content and performance standards envisioned by NCEST and *Goals 2000* go far beyond the current NAEP frameworks or item pools. The 1992 NAEP achievement-level descriptions are an intermediate hybrid, influenced by subject-matter experts with an appreciation for the conceptions of the future but also tied to the current tests.

If NAEP is to be used to monitor progress toward the national education goals, it is essential that NAEP content and performance standards be consistent with national standards. Coordination and compatibility with national standards must occur at several levels. Of primary importance, the content domains must be compatible.

However, in accord with the Panel's earlier recommendation that NAEP be comprehensive with respect to competing views of curriculum and current education practice, NAEP must have a more inclusive content domain than implied by future-oriented nationally certified content standards. For example, NAEP should continue to assess current curricular goals in addition to those implied by new content standards as a way to measure change over time. But NAEP cannot measure less than is implied by new content standards or it will be incapable of measuring progress toward their attainment.

NAGB, in its *Future of NAEP* report, has wisely charted a course to align NAEP with national content standards while at the same time maintaining the capacity to assess the outcomes of current practice.

> The Governing Board believes that appropriate alignment of the National Assessment with certified national standards is essential, that national standards should be a primary basis for developing assessments, that incorporation of such standards into the National Assessment should be done through successive adjustments of its frameworks and assessments, and that the goal should be to achieve a balance between the vision contained in new voluntary standards and the reality of current instruction.[8]

The Panel concurs with this policy about what should be included in the assessment. It should be emphasized, however, that *the goal should be to ensure that the content of NAEP is comprehensive, rather than to achieve a balance or compromise between competing perspectives.* A comprehensive assessment can be used to detect differential changes such as an increase in computational skills at the expense of problem-solving abilities. But as discussed in the next section, content that is a mixture of old and new summarized in a single score may not be adequate for measuring attainment of national standards.

NAEP must also be consistent with national standards at the level of performance standards. This means that *NAEP performance standards must be congruent with national performance standards and faithful to the substantive expectations underlying national content standards.* One of the reasons for having national standards is to convey to the public what good teaching and student learning should look like. Verbal descriptions, sample tasks, and scoring criteria reported along with assessment results are the basis for communicating concretely what the standards mean.

2. NAEP Must Remain a Comprehensive Assessment as Well as Integrate National Content Standards.

To reiterate, the Panel has emphasized that NAEP content should not swing wildly in response to each new curricular innovation but should be comprehensive, including both old and new conceptions of academic goals. The conclusion that NAEP frameworks must change to reflect national content standards means that NAEP will have to stretch to encompass new substantive expectations. Very likely, traditional content will become a smaller fraction of the total. But some fundamental incompatibilities between the old and new are also possible. Insufficient thought has

[8] National Assessment Governing Board, "Discussion Paper on the Future of the National Assessment of Educational Progress" (Washington, D.C.: Author, August 1992), 10.

been given to the decision to report assessment results reflecting different curricular perspectives as a single score. Nor has enough thought been given to how changes in this mixture over time will affect the interpretation of trends.

As national content standards are developed, it would be useful to ask what the NAEP assessment would have to look like to reflect the standards with no constraints from the NAEP of the past. Then ask what kind of assessment would have to be administered to maintain trend lines and to see how the nation is doing according to traditional goals and practice. Finally, ask whether these two assessments can reasonably be scored and interpreted as a single global measure. In recent years, NAEP frameworks have changed repeatedly, often with subject-matter experts straining against the need to preserve trend lines. The result has been an odd stew of old and new with neither perspective represented well by aggregate scores. An alternative is to administer the composite assessment but to report two pictures of progress. This approach is consistent with NAEP's current practice of reporting both old and new trend lines. (However, the substantive differences between a national-standards-aligned assessment and a current-practice assessment are likely to be greater than previous changes in NAEP frameworks.) The Goals 3 and 4 Technical Advisory Group suggested a similar approach because of their disquiet over current NAEP content. They recommended that "data be partitioned so that the reader can discern how students are doing on items aligned with the NCTM standards and those aligned with other standards."[9] Comparing NCTM versus traditional mathematics content or National Academy of Sciences/National Research Council science standards versus traditional science content has the greatest prospect for yielding useful information about changes in what students know and can do, just as NAEP data from the 1980s showed that basic skills were going up and higher-order thinking skills were going down.

3. Test Administration Procedures May Need to Change As the Assessment Incorporates National Content Standards.

As national content standards are certified, they will have a substantial impact on the NAEP frameworks and on the item pools as well (e.g., more performance-based items). Changes in the types and distribution of items will likely necessitate changes in test administration procedures. More extended problems may require more time per student or more replications. Furthermore, measuring new national standards will force NAEP to include more open-ended problems and performance assessments.

These changes, in turn, will have implications for implementation of the matrix sampling design as well as for the cost of administering NAEP. Perhaps more important is the impact on the individual test-taker. If serious efforts are made to measure advanced performance, using the customary administration procedures, how will a basic-level student respond upon receiving a booklet with a single 30-minute advanced problem? This would clearly be an unhappy experience for the student and, in addition, produce little useful data for the assessment. Accurate assessment of advanced performance when the new content standards are in place may require some kind of branched sampling of students, whereby students attempt different levels of problems based on a score on a placement or prescreening measure.

[9] Goals 3/4 Technical Advisory Group meeting, Council of Chief State School Officers (Washington, D.C.: April 29, 1993).

4. The Issue of Unidimensional Scales and Composite Scores May Have to Be Rethought.

Several findings from the Panel's studies call into question the appropriateness of relying so heavily upon composite-score reporting. The issue of distinct content standards within the larger assessment was addressed in an earlier section. In addition, content experts in mathematics have argued strenuously that composite-score reporting creates a misleading picture of subject-matter knowledge and instructional goals. Mathematics experts noted that in 4th grade, the predominance of the numbers and operations strand obscured other content areas. At 8th grade, differences in opportunity to learn algebra would create misleading composite results. Panelists at the 12th grade emphasized that only by reporting by each content strand separately would it be possible to communicate to the public that students need to perform acceptably in all of the areas.

Most seriously, in trying to wrestle with conceptions of proficient and advanced performance, findings from several different sources suggest that a single composite scale cannot adequately represent the way that knowledge and abilities are acquired. A plausible explanation for wildly discrepant findings between item judgments and borderline exemplar cutpoints is that judges mistakenly expected students' given performance (on one open-ended problem) to be a perfectly reliable indicator of their performance on all such tasks. A similar expectation may explain why the expert panels in reading and mathematics using 80-percent-correct item maps set such high cutpoints. In contrast, in the whole-booklet study and in the field studies, judges set lower cutpoints (for the same narrative standards) when they saw whole booklets or other more complete demonstrations of student knowledge and abilities.

Trying to locate proficiencies along a single continuum assumes that knowledge acquisition works like filling a measuring cup. If the liquid rises to a certain level, then it is assumed that everything below that point is full or known and everything above that point is empty and unknown. As common sense might suggest, the NAEP scales do not work this way. Proficient performance is more often a combination of some basic-level skills, some evidence of abilities and performance we call proficient, and even some evidence of advanced knowledge. Similarly, students whose total performance is persuasively advanced when viewed in its entirety still miss some basic or proficient items.

Additionally, separate scores on substantive subscales would facilitate the interpretation of achievement as a function of opportunity to learn, a central concern in the pending *Goals 2000* legislation. They would also provide the public and practitioners with more information on the relationship of achievement to instructional practice.

The Panel strongly suggests that the practice of relying so heavily on a single composite NAEP score scale in each subject area be seriously re-evaluated. Action should be taken immediately to give greater attention to the further creation of subject-matter subscales. In addition, as the development of performance standards congruent with national content standards is undertaken, the fit between substantive conceptions of proficient and advanced work and statistical score reporting should be rigorously examined.

Final Observations of the Panel

The evaluation of the 1992 NAGB achievement levels has been an odyssey of discovery. Much has been learned, not only about the procedures and results of the NAGB effort, but about the establishment of standards for American education. The task NAGB attempted was unprecedented. Never before had there been any attempt to establish cutscores representing high and rigorous education performance standards in such a complex, pressure-laden technical and policy context. The work has been a stimulus to the creation of significant new knowledge, both technical and substantive, concerning the issues that must be confronted in articulating content standards, assessment frameworks, verbal descriptions of performance standards, numerical cutscores, and item exemplars. At the same time, the work has assured opportunities for participation in that process by a range of constituencies from policymakers to content experts to qualified representatives of the general public. No one could have imagined either the range of difficulties encountered or the range of evaluation studies that have been developed as a consequence of NAGB's initiative. Together with the many other psychometricians, curriculum experts, education policy specialists, and other researchers and scholars who participated in evaluating earlier NAGB achievement levels, members of this NAE Panel have learned a great deal about the processes by which achievement levels may be established and validated.

The members of the NAE Panel strongly affirm the potential value of performance standards and test-score achievement levels when linked to comprehensive content standards as a way to establish clear expectations for teachers, students, and the public. Furthermore, we strongly affirm the value of performance standards as one method for NAEP reporting, both to improve the interpretability and usefulness of the Nation's Report Card and to provide information to further the nation's education improvement. NAGB endeavored to improve the ability to interpret NAEP and used an established methodology to determine achievement levels. Nonetheless, as summarized in this chapter, it is the Panel's conclusion that the 1992 NAEP achievement levels should not be used.

The NAGB effort was credible, but the available technology was simply inadequate to the task. The design and implementation of the modified Angoff process was imperfect, but poor implementation was not the fundamental problem. Indeed, there may never have been a more elaborate (and certainly never a more carefully scrutinized) application of Angoff standard setting. Rather, the Panel concludes that it is an impossible cognitive task to make accurately the kind of judgments called for in using an Angoff procedure in this context. Even with the most careful training and the best of intentions, human judges apparently cannot integrate (1) a conception of a population of examinees at the borderline between two ranges of performance; (2) an inference about the knowledge and skills elicited by a given item; and (3) the effects of the item's format, quality and character of distractors, probability of a correct response by guessing, and other factors in order to arrive at a reasoned conclusion about the probability that a borderline examinee will answer that item correctly. It is not just that such judgments have a large component of random error. If it were simply that, then the problem could easily be solved by increasing the number of judges and the number of items. Rather, the Panel finds the evidence compelling that such judgments are subject to large biases, which cannot be adjusted for because they cannot be estimated by any known procedure.

There is an appearance of objectivity, even scientific rigor, in any procedure that involves so many numbers and abstruse calculations, but it is not science. The standards that result are not defensible. It would have been far better, in the Panel's view, if NAGB had simply chosen several base-year percentiles to serve as benchmarks (as described earlier in our short term recommendations) and then relied on their collective judgment to arrive at targets for the proportions of students expected to reach those levels in future years. In addition to saving the very considerable expense of the procedure used, a benchmark-percentile approach would have resulted in achievement levels with no unwarranted substantive assertions or implications attached to them; and that almost surely would have been plausible and coherent in the sense of showing a logical progression across grade levels. In contrast, the levels arrived at by the modified Angoff procedure invite unwarranted interpretations about the actual capabilities of students at specified levels. Any plausibility and coherence of the proportions they imply are largely matters of happenstance, helped along by feedback between rounds of rating that was designed to nudge judges in the desired direction without acknowledging the essentially normative character of the exercise.

The Panel believes that NAEP itself should not function directly as an agent of education reform. If NAEP policy decisions are made to "leverage" education change, there is the risk that the indicator itself will become corrupted. If NAEP actively pursues a reform agenda, the content of the indicator will be dictated by the reform curriculum advocated at a given time and will shift erratically over time in response to changes in political authority. The Panel believes, however, that NAEP has an important role to play in education reform by providing data that inform the debate about needed changes. The Panel also affirms the desirability of voluntary national standards developed by agencies other than NCES or NAEP as a tool to lead education change and reshape public thinking about the substance and level of academic expectations. Furthermore, to be able to report on progress regarding student attainment of national performance standards linked to national content standards, it is essential that NAEP mirror national standards.

Earlier in this chapter, the Panel suggested some essential elements for what it believes would be an acceptable procedure for establishing performance standards. The Panel is not simply suggesting its own preferred method in opposition to the modified Angoff procedure. That was not the charge. The Panel believes that a defensible procedure is well within reach, however, largely due to the pioneering efforts of NAGB, its contractors, and the many evaluators of the 1990 and 1992 NAEP assessments. The Panel looks forward to the promulgation of rigorous and defensible achievement levels for NAEP, but cautions that it may take some time to establish them. To assist in reaching that objective, the Panel has recommended criteria and procedures for improving the interpretability and usefulness of NAEP reporting, for grounding NAEP in emerging national content standards, and for assuring the continued credibility of NAEP as an essential indicator of achievement in American education.

Appendix

Synopses of Studies for the National Academy of Education Panel on the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment: An Evaluation of the 1992 Achievement Levels

This appendix contains summaries of 10 studies that were commissioned by the Panel and served as the basis for the Panel's findings and recommendations. The complete reports are being published in a separate volume, *Setting Performance Standards for Student Achievement: Background Studies*.

An Evaluation of the 1992 NAEP Reading Achievement Levels, Report One: A Commentary on the Process

David Pearson and Lizanne DeStefano
University of Illinois at Urbana-Champaign

In an era of educational reform characterized by accountability, assessment, and public awareness, the National Assessment of Educational Progress (NAEP) has been asked to change in response to the needs of practitioners, the public, and policymakers. By Congressional mandate (P.L. 100-297), the National Assessment Governing Board (NAGB) was charged with the task of identifying achievement goals for each grade level in each subject area tested by NAEP. Beginning in May 1990, NAGB chose to address this mandate by seeking to define achievement-level standards of basic, proficient, and advanced student performance in terms of scores on the NAEP performance scale for each subject area at each grade level and to report results using these standards.

For this study, the investigators focused on the process followed in setting the standards for the field of reading. Specifically, they sought to answer two questions:

> What was the nature of the level-setting process for reading? Was it technically sound and professionally credible? What were its strengths and weaknesses?
>
> To what extent was the level-setting process grounded in the NAEP Reading Framework? To what extent was it consistent with the issues and structures presented in the Reading Framework?

The evaluation approach used for this study combined both qualitative and quantitative techniques. The six-member research team observed general and grade-level sessions at the St. Louis meeting in August 1992, when achievement levels for reading were set through the use of a modified Angoff process. The researchers also conducted group and individual interviews with participants, staff, and other observers and collected materials, products, and other hard-copy items. After preliminary analyses of these data, the research team developed a followup survey that focused on four issues: (1) participants' perceptions of their role and the purpose of the meeting; (2) the referent group used in the development of descriptors and item ratings; (3) the individual process of rating and confidence in final ratings; and (4) personal definitions of reading. In addition, two members of the research team observed the achievement-level-descriptor review in San Diego in October 1992, gathering data that paralleled the St. Louis data collection. A focus-group session was also held, with experts reviewing the descriptors, and observations were made at an Educational Testing Service meeting to develop descriptors using anchoring and item-mapping data.

The Level-Setting Process as Intended

To frame their evaluation of the level-setting process as observed in St. Louis and San Diego and described by participants and planners, the investigators stated the assumptions, both implicit and explicit, that should hold true if the process is to be judged technically sound and credible. The assumptions are outlined below.

1. Through the creation of ad-hoc advisory panels and the involvement of a heterogeneous group of judges (teachers, nonteacher educators, and representatives of the general public) in level setting, not only will the level-setting process be accessible to a broad range of professional and lay people, but the process will be improved when a panel of judges includes individuals from a broad spectrum of society.

2. Because the Reading Framework was identified as the basis of the development of the achievement-level descriptors and the descriptors were intended to represent student performance on the assessment, a strong relationship exists between the framework and the 1992 NAEP in reading and subsequent assessments.

3. The framework serves as a link between the NAEP in reading and the descriptions of the achievement levels; only in this way can the descriptors remain stable even though the items might change from one assessment to the next. It was assumed that participants were familiar with the framework and used it as *a* primary basis, if not *the* primary basis, of the descriptors they developed.

4. When individuals rated items, they were judging in terms of a common referent group, and the descriptors (and therefore the Reading Framework) played a primary role in helping them identify the hypothetical groups that existed at each cutpoint.

5. During the rounds of the modified Angoff process, feedback about actual student performance of judges improved the accuracy of the item ratings and the validity of the entire process.

6. In the scoring of constructed-response items and the selection of exemplars, actual student responses represented the full range of borderline performance across all three achievement levels.

The Level-Setting Process as Observed

The investigators determined the extent to which their data supported or refuted the validity of each assumption and used their results to answer the two study questions listed above.

Regarding the first study question, the investigators believe it is fair to conclude that American College Testing (ACT) carried out the level-setting process as proposed; in fact, they went to great ends to preserve the integrity of the modified Angoff process used and to adhere to strict time schedules. However, when the process was applied to a situation which involves heterogeneous judges, multiple cutpoints, and a complex domain of items (multiple choice, short answer, and constructed response), three key assumptions (1, 5, and 6) were violated.

Regarding Assumption 1, the investigators found that the level-setting process was not accessible to a broad range of professional and lay people; in fact, the inclusion of a wide range of judges created problems with the development of the descriptors and the consistency of the reference group. In violation of Assumption 5, the investigators found that feedback about actual student performance and the relative performance of judges, in the absence of substantive discussion and negotiation, did not lead to consistent item ratings. In regard to Assumption 6, the investigators determined that actual student responses did not represent the full range of borderline performance across all three achievement levels. Representation was most lacking for the constructed-response items at advanced achievement levels.

In light of these findings, the investigators see major problems in the technical soundness and professional credibility of the modified Angoff process applied in this situation. While ACT carried out the process as proposed, the process itself was fundamentally flawed. With respect to the second question, focusing on the relationship between the framework and the level-setting process, the investigators found little evidence to support any significant relationship between the two.

Specifically, Assumption 2 was violated: the relationship between the Reading Framework and the 1992 NAEP in reading was not as strong as intended or as represented to participants. This created problems when trying to develop descriptors that represented both the framework and the assessment. In reference to Assumption 3, it was found that the participants in St. Louis were not familiar with the framework, and they did not use it as a basis for the descriptors, preferring to use personal referents of the performance of actual students. The San Diego participants made greater use of the framework when revising the descriptors but were constrained by the fact that they could not make substantive changes in the St. Louis descriptors without compromising the level-setting process. The final version of the descriptors submitted for approval by NAGB reflects the framework more consistently. At a surface level, it is substantially different from the version available in St. Louis. However, a close analysis by the panel of experts suggested that even the final version fails to capture the constructive epistemological intentions of the framework. In short, despite the rhetoric, it is still shrouded in a set of skills-based assumptions about reading. Finally, with respect to Assumption 4, in rating items, it was found that judges tended to use the performance of actual referent groups from their own experience, rather than hypothetical groups defined by the descriptors, when assigning ratings to items.

Given these findings, the investigators cannot support achievement-level reporting of the 1992 NAEP in reading. Their only question is whether the failure resides in this particular instantiation of the modified Angoff process or is inherent in the process itself. Clearly, the investigators have evidence for the first interpretation. All six of the assumptions underlying its application to the 1992 NAEP in reading were violated. Regarding the second interpretation, to the extent that the six assumptions are particular to this application, the Angoff method may be salvageable. However, the investigators believe that only Assumptions 2 and 3 are unique to 1992 NAEP in

reading. Assumptions 1, 4, 5, and 6 seem relevant to any application of the process and therefore bring the process itself into question. It may be that the demand characteristics of a process that requires the assignment of ratings by imagining the *actual* performance of a *hypothetical* group based on *ideal* achievement is beyond the cognitive capacity of even the most knowledgeable judges.

For that reason, the investigators recommend that NAGB invest in a fresh approach of the standard-setting process, taking care to avoid the severe limitations of the modified Angoff approach identified in these evaluation studies. Only in this way can the 1992 assessment become a valid baseline for anchoring future trend analyses.

Validity of the 1992 NAEP Achievement-Level-Setting Process

Donald H. McLaughlin
American Institutes for Research

Beginning in May 1990, the National Assessment Governing Board (NAGB) implemented a new format for reporting National Assessment of Educational Progress (NAEP) results—achievement levels, which are statements of how basic, proficient, and advanced students at grades 4, 8, and 12 should perform. The process was repeated and expanded in 1992. The panelists who participated in the 1992 standard-setting process produced large numbers of numerical ratings and made quantifiable selections, and the purpose of the investigator's study was to address questions about the validity of the process used in 1992. Specifically, the study addressed the question of whether the assumptions underlying the analysis of the judgment data gathered in generating the 1992 NAEP achievement levels were valid, based on internal analyses of the data. The three broad assumptions examined were:

1. Did each panelist generate responses based on a clear conception of borderline basic, proficient, and advanced performance?

2. Was there round-to-round improvement in the consistency of the panelists' responses?

3. Did the panelists reach a consensus by the third round of ratings?

With data used by American College Testing (ACT) to generate cutpoint estimates, the investigator conducted analyses that paralleled those conducted by ACT. Because of the difference in objectives (production versus evaluation), Dr. McLaughlin conducted somewhat different analyses from those ACT conducted, using a method of cutpoint estimation that provided the flexibility required to measure variations in cutpoint estimates. All analyses were parallel for reading and mathematics, and a measure of the reliability of the results of the evaluation can be inferred from the similarity of the findings for the two content areas. The results fall into three categories that relate to the assumptions above; each category is summarized below.

1. Validity of the Cutpoint Estimates

The four comparisons described below examined the assumptions underlying the establishment of cutpoints for minimal basic, proficient, and advanced achievement levels. The first comparison is between items and judgment methods, and the last three are between items, holding the judgment method constant.

Dichotomous Items Versus Extended-Response Items. The first assumption tested was whether the Angoff judgmental process, when used to determine achievement levels for dichotomous ("right/wrong") items, yields the same cutpoints as the Boundary Exemplars judgmental process, when used to determine achievement levels for extended-response items. To use the Angoff procedure, panelists estimated the percentage of students at the threshold of each level who would "get the item right." For the Boundary Exemplars procedure, panelists were given several dozen examples of actual responses and were asked to identify the ones most indicative of threshold performance at the three levels.

The estimates obtained for the two item types differed by more than 50 points for mathematics, and the differences for reading were even larger. These differences between the cutpoints based on the two item types in mathematics are as much as the growth across four grades, and in reading the differences are as much as the growth across eight grades. The investigator concluded that the assumption that panelists were imagining a single level of performance at each cutpoint and making the judgments in line with that level of performance for both item types must be rejected, along with numerical results based on that assumption.

Variations in Cutpoints Between Multiple-Choice and Short-Answer Items. The Angoff procedure assumes that panelists' judgments will remain consistent, across different subsets of the NAEP items. The investigator compared the differences between multiple-choice and short-answer open-ended items. Panelists had not been given instructions regarding whether and how to take guessing into account, and it was expected that their responses might reflect the number "who know and can do the problem," as opposed to the proportion who might get it right, including by guessing. In such a case, panelists would generate smaller numbers for their responses to multiple-choice items than were warranted, leading to lower estimated cutpoints for the multiple-choice items.

The investigator found that this was a substantial effect, especially for cutpoints at the basic level. Many panelists apparently did not take guessing into account, and aggregate results are somewhat biased by this factor. The direction of the bias is not clear because there has been no policy discussion as to the appropriate weight to assign to skills assessed by multiple-choice versus short-answer items. The investigator considered the result to be a serious deficiency in the validity of the judgment process.

Variations in Cutpoints Between Easy and Hard Items. The NAEP assessment, like all such assessments, includes some items that are easier than others: more students are expected to know the answers to some items than to others. The Angoff procedure assumes that panelists can take into account these differences in item difficulty, estimating lower percents correct for difficult items and higher percents correct for easier items. If panelists fail to take difficulties into account, the estimates for easy items will be substantially lower than warranted.

In comparing the achievement-level cutpoints, the investigator pooled first within multiple-choice items and then within short-answer items, and he compared the easiest half of each item type and the hardest half of each item type. The results for both mathematics and reading indicated that panelists were not adequately taking into account the differences between easy items and hard items. This was true both before and after panelists were shown the item p-values. The investigator concluded that the task required skills in item interpretation that were beyond the reach of panelists, in the context of the training and instructions given and the time allotted.

Variation in Cutpoints Between Subscales. The NAEP instruments were based on a framework containing five mathematics subscales and four reading subscales, and because the achievement levels report a single composite for mathematics and reading, cutpoints were obtained by combining cutpoints across subscales, using weights determined by policy. To determine whether panelists considered typical student performance in some subdomains to indicate greater achievement than performance in others, the investigator compared the cutpoints by subscale.

The differences found were generally not large. The only exceptions were for eighth-grade mathematics, for which somewhat higher advanced standards were set for measurement and for data analysis, statistics, and probability than for other subscales. The investigator compared cutpoints for all items and separately for dichotomous items and found no tendency for either method to show more variation between subscales than the other.

2. Changes Between Rounds

The standard-setting task involved a series of three rounds of ratings of the same items by panelists, and the idealized model assumes that a process of accommodation among the panelists, working together, will lead to a consensus accepted by all participants. The multiple rounds are also intended to provide an opportunity to assess panelists' judgments both before and after they receive information about the actual performance of the population on the items. Changes in rounds either can affect the mean level of cutpoints estimates or can increase consistency and consensus without changing the cutpoints.

The investigator examined the changes in cutpoints by round and found that the shifts in mean levels were generally small, with only modest reduction in variance among raters. Panelists did use the information provided between Rounds 2 and 3 to improve the match between their estimated "percents correct" and the difficulty of the items. However, the results of the investigator's other analyses suggest that even in the final round, panelists were not generating ratings consistent with the assumption that they had a clear concept of borderline basic, proficient, and advanced performance.

3. Variance of the Cutpoint Estimates Between Panelists

A major assumption of the judgment methodology is that a consensus among qualified panelists can be developed, across specified demographic groups and among all panelists. To test this assumption, the investigator examined cutpoint variations (in the

final round) both overall and by demographic groups (teachers, nonteacher educators, members of the public; whites and all other races; men and women).

None of the differences found between groups was statistically significant. A parallel analysis of group differences in average internal consistency was also carried out, and no significant patterns were found. Nevertheless, the variance between individual raters was large, even in the final round, and it cannot be concluded that by the final round there was sufficient improvement to ensure the validity of the ratings.

Conclusions

Based on these results, the investigator concluded that the procedure employed to generate achievement-level cutpoints on the NAEP scales in mathematics and reading in 1992 was seriously flawed. The panelists were called upon to complete a task that was apparently too difficult. A fundamental question must be addressed: *Can panelists be found and prepared who can validly conceive of NAEP achievement, and, moreover, validly estimate the percent of students at an achievement level who would get a test item right?* If not, then the Angoff procedure must be abandoned for this task.

Perhaps the greatest flaws in the 1992 procedure occurred in the use of the Boundary Exemplars method for identification of cutpoints on extended responses. Before this method is used in this context again, a design that avoids bias due to the range of exemplars provided must be implemented, and a method of analysis that accurately translates from selections to NAEP scale scores must be implemented. The sensitivity of the achievement levels to the proportions of dichotomous and extended-response items on the assessment, to the proportions of multiple-choice and short-answer questions, and to the distribution of hard and easy items, limits the usefulness of the levels for reporting an evolving NAEP assessment in the future.

Order of Angoff Ratings in Setting Multiple Simultaneous Standards

Donald H. McLaughlin
American Institutes for Research

The 1992 standard-setting process for the National Assessment of Educational Progress (NAEP) broke new ground in many ways and has brought to the attention of the education community a variety of issues with respect to the interaction of curricula, teaching, standard setting, and assessment construction. Briefly, the 1992 NAEP achievement-level-setting process brought together, for each of three grades (4, 8, and 12) and subject matters (reading, mathematics, and writing), a single panel of teachers, nonteacher educators, and qualified members of the general public for a 3- to 4-day

meeting in St. Louis. After becoming familiar with NAEP, the panelists settled on operational definitions for *basic, proficient,* and *advanced* performance at their grade in their content area.

This empirical study is part of the National Academy of Education (NAE) validation of the 1992 NAEP standard-setting process. In setting the 1992 standards, NAEP conceived of performance in an area such as mathematics or reading as complex and multidimensional, but assessable in terms of a small number of domains, or scales, which in turn could be summarized as a single composite. In this way, basic, proficient, or advanced performance in an area could be represented by NAEP assessment scores above specified "achievement-level" cutpoints.

The achievement-level cutpoints for mathematics and reading that resulted from the 1992 NAEP standard-setting process followed a very regular pattern. For mathematics, at each grade the difference between basic and proficient cutpoints was about 40 NAEP scale-score points (39 for grade 4, 41 for grade 8, and 43 for grade 12) and between proficient and advanced was about 35 points (32 for grades 4 and 12 and 36 for grade 8). For reading, the basic-to-proficient cutpoint differences were about 35 (31 for grades 4 and 12 and 38 for grade 8), and the proficient-to-advanced intervals were in the same range (32 for grade 4, 36 for grade 8, and 44 for grade 12). This pattern has been interpreted as support for the reliability and validity of the standard-setting process.

The Angoff Procedure Examined

To interpret the regularity of the pattern of achievement-level cutpoints, it is essential to eliminate possible artifactual explanations, one of which is associated with a technical innovation incorporated into this application of the Angoff procedure. The innovation was the request for panelists to estimate simultaneously three different achievement levels in one domain: basic, proficient, and advanced. This task requires each panelist to imagine, successively, a student who is at the cutpoint for basic, a student who is at the cutpoint for proficient, and a student who is at the cutpoint for advanced. Because the instructions were to make all three estimates for each item before proceeding to the next item, a panelist was required to switch cutpoints twice for each item.

Panelists may not have made the three estimates for an item independently: they may have made a single estimate based on the image of a student at one cutpoint and then added (or subtracted) what they believed to be an appropriate amount for the other two cutpoints. If this were true, the regularity of the pattern of cutpoint estimates would be overstated. The purpose of this study, then, was to address this question:

> To what extent is the regular pattern of differences among achievement levels generated for NAEP in 1992 a function of this ordering of the panelists' task?

Having decided to carry out this study, the NAE panel staff also conducted analyses of the study data to address two other important questions:

Was there evidence that changing the verbal descriptions of the levels after the completion of the rating process, as was done in the 1992 NAEP achievement-level-setting process, would lead to a different setting of cutpoints?

How robust would the achievement levels be across two quite different panelist-sampling processes and two different sets of instructions and training procedures?

The first of these two questions is of critical importance for the validity of the reports of 1992 NAEP mathematics data in terms of verbally described achievement levels; however, the sample size of the study limits the precision of the test. Results of analyses to address the second of the two questions should be interpreted with great caution because if differences are found, they may be due to many of the differences between the two Angoff implementations.

Methods

This small-scale empirical replication of the standard-setting procedure used by NAGB was performed as part of another study carried out for NAE by the investigator (see "Rated Achievement Levels of Completed NAEP Mathematics Booklets," also summarized in this volume). From nominations obtained from school district superintendents in the San Francisco Bay Area, 24 individuals (13 teachers, 4 nonteacher educators, and 7 members of the general public), who were familiar with eighth-grade mathematics performance or interested in it, were selected. During the day prior to participation in this study, they had participated in the investigator's other study; hence they were familiar with NAEP and performance of eighth graders on NAEP. During the all-day session, panelists were assigned to specific seats at one of four tables, each with representatives of the teacher, nonteacher educator, and general public groups. Assignment to tables and to experimental conditions was random within strata of teachers/nonteacher educators/public and by male/female and by white/all other races.

Materials were 50 items used in the NAEP 1992 eighth-grade mathematics assessment. They were selected from four blocks of items that did not appear in the two NAEP item booklets that the panelists had examined during the preceding day. The items were prepared in stapled booklets, with different blocks on different pages. Right-wrong scoring rubrics were provided for open-ended response items. Panelists were instructed in the use of the Angoff procedure and given a trial Angoff task. With one of two types of response sheets, panelists were asked to record the item percents correct: half generated the Angoff ratings in the same manner as in the NAEP 1992 standard-setting process (generating all three cutpoint ratings for one item before proceeding to the next item), and half generated the ratings in the "opposite" manner (i.e., rating all items on one of the cutpoints before proceeding to the next of the three cutpoints). Panelists were aware of the differences between the two forms of the response and the general purpose of the comparison, but they did not know which version corresponded to the procedure used in St. Louis. At the end of the day, panelists were asked to complete a questionnaire asking them how they performed

the task and what they thought of the various stimulus materials, including the achievement-level descriptions. An independent observer took notes on the progress of the study and submitted a report.

Results

The major finding from this study is that the use of the three-ratings-per-item procedure did not introduce substantial and statistically significant artifactual regularity into the cutpoints. Nonsignificant differences in the predicted direction were found, but the study was designed so that a difference that would be considered important, such as a halving of the standard deviation of differences due to the order of presentation, would be likely to be found to be statistically significant. Therefore, it does not seem warranted, based on these results, to recommend against the use of the three-ratings-per-item methodology.

A secondary finding, observed but not apparent in the data presented here, is that the method of rating all items at one level before proceeding to the next level did not take longer than the method of generating all three ratings for an item at one time. The order of completion of the forms by panelists was recorded, and there was no tendency for panelists in one group to take longer than panelists in the other group. Therefore, if effort is the major concern that might lead to use of the three-ratings-per-item method, that concern was not borne out in this study.

This study also offered an opportunity for two additional comparisons. First the comparison of cutpoints generated between the original (St. Louis) version of the achievement-level descriptions and the revised (Nantucket) descriptions indicated no significant differences at the basic and advanced levels and only a marginally significantly higher proficient cutpoint by panelists using the Nantucket description. The investigator does not believe that the level of discrepancy observed in this study is cause, in itself, for judging the revisions made in Nantucket to have invalidated the achievement levels.

Finally, this study was a very rough replication of a portion of the 1992 NAEP achievement-level-setting process, and given the substantial, intentional variations between the procedures, the results for the proficient and advanced levels were remarkable similar, while the differences observed for the basic level can be attributed to a particular variation in the training that would affect the setting of cutpoints in the lower part of the score distributions.

Rated Achievement Levels of Completed NAEP Mathematics Booklets

Donald H. McLaughlin
American Institutes for Research

The National Assessment Governing Board (NAGB) sponsored an achievement-level-setting effort in order to be able to present the results of the 1992 National Assessment of Educational Progress (NAEP) as percentages of the nation's students who show evidence of performing at the basic, proficient, and advanced levels. To set the achievement levels for grades 4, 8, and 12 in reading, mathematics, and writing, panels of approximately 12 teachers, 4 nonteacher educators, and 6 qualified members of the general public (one panel per grade level and content area) met for 3 to 4 days in St. Louis. After agreeing on an operational definition for "basic," "proficient," and "advanced" performance at their grade in their content area, they were asked to make judgments about the specific item-level performance of students at the (lower) threshold of each achievement level. For "right/wrong" items (multiple choice and short answer), the Angoff procedure was used to gather judgmental data, with panelists estimating the percentages of students at the threshold of each level who would "get the item right." For extended-response items, the Boundary Exemplars procedure was used, with panelists reviewing examples of actual responses and identifying those most indicative of threshold performance at the three levels. Panelists repeated the rating process for three rounds. Results were compared to the item-response-theory (IRT) parameters of the items to determine the NAEP scale points that best reflect the basic, proficient, and advanced ratings provided at each grade level. Later, after the St. Louis sessions, the descriptions of the levels were revised by another group of panelists to reflect more closely the state-of-the-art frameworks for curriculum in the different areas, and the revised descriptions became the official versions for reporting purposes.

The two primary products of the 1992 achievement-level-setting sessions were: (1) verbal descriptions, or characterizations, of basic, proficient, and advanced performance and (2) NAEP scale-score cutpoints between the levels. The descriptions and cutpoints were used in the reports of the 1992 assessments. It is important to know whether the audiences for the NAEP reports—teachers, nonteacher educators, and members of the general public having interest in education—will interpret the words *proficient, basic,* and *advanced* used in those reports in the same manner that the panelists who set the cutpoints did. If the panelists held a concept of "advanced" as a rare, elite performance, for example, while readers of the reports considered "advanced" to refer to performance levels that are achieved by the top 10 percent of students, then NAEP reports would be misleading. This study addresses that concern.

Purpose of the Study

The study focused on two questions:

1. Would a group of individuals similar to those who set the 1992 NAEP achievement levels, using the verbal descriptions of the levels, place booklets of NAEP performance in categories (advanced, proficient, basic, and below basic) that match the categories of the NAEP scale scores assigned to those booklets?

2. Would individuals similar to those who set the achievement-level cutpoints categorize booklets of NAEP performance differently if they used the revised verbal descriptions, as opposed to the original verbal descriptions (i.e., those developed in St. Louis)?

In the interests of economy, the study was carried out only for eighth-grade mathematics, with no prior expectations for the outcome. The decision was made to review the study results before considering its extension to other grades and areas. If results indicated that study participants interpreted the achievement-level descriptions very differently from the panelists in St. Louis, recommendations of caution in interpreting achievement levels for other grades and subjects would be appropriate.

Study Methods and Procedures

Nominations were solicited from school-district superintendents in the San Francisco Bay Area for individuals who were familiar with eighth-grade mathematics performance or interested in it, and 24 individuals (13 teachers, 4 nonteacher educators, and 7 members of the general public) were selected. During an all-day session, panelists were assigned to one of four tables, each with representatives of the teacher, nonteacher educator, and general public groups.

For the study, 160 NAEP booklets with responses from eighth graders in 1992 were used, and three copies of each were made, for a total of 480. Each panelist was given a packet of 20 booklets, sorted with the highest scoring booklet on top and the lowest on the bottom. Half of the 160 booklets had one set of items; half had another set. However, each panelist compared the performances of 20 students given the same set of items. The 160 booklets included advanced, proficient, basic, and below basic performances, but booklets were assigned so that each panelist would see different distributions of performance. For example, one group of packets included two booklets completed by students classified as advanced, eight proficient, four basic, and six below basic; the other three groups of packets had other combinations. Classification was based on a "normit percent correct" computed by the Educational Testing Service because scale scores were not available at the time booklet selections were made.

Panelists at two tables were given the achievement-level descriptions developed by panelists in St. Louis who generated the Angoff ratings used the determine the 1992 NAEP mathematics achievement-level cutpoints. Panelists at the other two tables were

given the descriptions as revised at the conference in Nantucket several weeks after the St. Louis meeting. All 24 panelists were instructed to use the definitions supplied and to classify the performance of each student as advanced, proficient, basic, or below basic, placing each booklet into one of four categories corresponding to those levels. Panelists were told that there were eight different packets, with no two people at the same table having the same booklets, and that they might have no advanced students' booklets or many. They were told to compare each booklet to the descriptions of the achievement levels, to determine the best place for each booklet. They were told that even though the booklets were in order, they could question the ordering, and they were given a specific example of a rationale for ignoring the ordering.

Results of the Comparison of Cutpoints

The primary objective was the comparison of NAEP cutpoints with the boundaries between the categories generated by the study participants. Data were analyzed to identify the best-fitting three cutpoints for each panelist and to compute the overall result as an average of the panelists' cutpoints.

The cutpoints estimated for this study were 253, 297, and 319. A comparison of the achievement-level cutpoints for the "basic" and "proficient" levels obtained by the two methodologies—whole-book categorization and the NAGB-sponsored item judgments—indicated agreement. A significant difference was found at the "advanced" level: the study panelists categorized 74 of the 480 booklets as advanced, but NAEP scale scores would have assigned only 21 of the booklets to the advanced level. In comparing cutpoints generated by the St. Louis version of the achievement-level descriptions and those that used the later official descriptions, the investigator found that at all three levels, the average of the assigned cutpoints based on the St. Louis descriptions were higher, although the individual results were not statistically significant.

Concerns that achievement-level cutpoints reflecting consideration of whole booklets by members of the potential audience for NAEP reports might differ by 20 or 30 points from the official values were not borne out. The statistically significant difference found at the advanced level, however, should be taken into account in considering the results of other components of the validation.

The panelists for this study participated in a companion study on the following day, using the item-based Angoff procedure, similar to that used in St. Louis. The overall cutpoints generated in that study were 224, 298, and 340, values that are more extreme than those generated by the whole-book rating process.

The two studies in combination did not exhibit any systematic bias due to the revision of the descriptions. The comparisons between the panelists using the St. Louis descriptions and the panelists using the final descriptions were, in fact, reversed for the two studies. The same panelists who manifested higher standards for the St. Louis descriptions (than for the official descriptions) in the study described in this summary manifested lower standards using the Angoff method than their colleagues who were given the official version. The investigator notes that this reversal is not surprising in that the differences are of a size to be expected by chance (that is, not statistically significant).

An Evaluation of the 1992 NAEP Reading Achievement Levels, Report Two: An Analysis of the Achievement-Level Descriptions

David Pearson and Lizanne DeStefano
University of Illinois at Urbana-Champaign

The second in a series of three reading studies, this investigation examined the quality of the achievement-level descriptions associated with the 1992 National Assessment of Educational Progress (NAEP) in reading. It included the review and critique of three versions of the achievement-level descriptions, listed in the order in which they were developed:

- ***the St. Louis descriptions,*** developed by participants in the St. Louis level-setting sessions and intended for use in the level-setting process
- ***the San Diego descriptions,*** a refinement of the St. Louis descriptions developed by an expert panel at a meeting in San Diego
- ***the official descriptions*** (the final version), a refinement of the San Diego descriptions carried out by a group of consultants and adopted by the National Assessment Governing Board (NAGB) for use in reporting the results of the 1992 NAEP in reading.

The study addressed three primary questions and several related subquestions:

1. What is the quality of the St. Louis descriptions as a basis for level setting? Are the descriptions consistent with the conceptualization of reading presented in the Reading Framework? Are the descriptions consistent across achievement levels? across grades?

2. Are there substantial differences between the St. Louis descriptions and the official descriptions? If so, are these differences enough to invalidate reporting by achievement levels for 1992?

3. What is the quality of the official descriptions as a basis for reporting the results of the 1992 NAEP in reading? Are the descriptions consistent with the conceptualization of reading presented in the Reading Framework? Are the descriptions consistent across achievement levels? across grades?

As in their first study, the investigators selected an evaluation approach that combined both qualitative and quantitative techniques, optimized the collection and triangulation of data from multiple sources, and focused upon faithfully representing the perceptions and recollections of experts in the field of reading and of various stakeholders involved in the level-setting process. The study included seven evaluation activities; each is summarized below.

Observations of the St. Louis meeting, held in August 1992 to develop descriptions that could serve as the basis for setting levels on NAEP, were made, and data were collected from grade-level and cross-grade-level observations, interviews, and document reviews.[1]

A followup survey was conducted to address participants' understanding of the purpose of the meeting, ideas about reading, referent group used in assigning ratings, individual process for rating, confidence in the descriptions and ratings, and general impressions.

Observations of the San Diego meeting, held in October 1992, were also made, and data were collected from grade-level and cross-grade-level observations, interviews, and document reviews.

Followup phone interviews were conducted with the San Diego participants, addressing their understanding of the purpose, ideas about reading, referent group used in developing descriptions, confidence in descriptions, and general impressions.

An expert panel was convened with three individuals with expertise in the areas of reading and literacy, assessment, and reading assessment and with familiarity with the Reading Framework and NAEP in reading. Four questions framed their discussions (e.g., Are the descriptions consistent with the conceptualization of reading presented in the framework?), and each participant wrote summary statements with their responses to the questions. Later, when the official descriptions were available, they were sent to the experts for reactions and responses.

An alternate level-setting session was held; participants were educators who were familiar with both grade-level performance and test development.[2] At part of this session, participants were asked to read and discuss the official version of the descriptions and to comment on them in writing.

Teacher interviews were conducted as part of the 1993 spring field test for NAEP in reading. Approximately 60 fourth- and eighth-grade teachers were asked to classify students in their classrooms into the categories defined by the official achievement-

[1] This activity is described more fully in the investigators' first study for this series; see their "Report One: A Commentary on the Process," also summarized in this volume.

[2] This session is reported in the investigators' third study for this series; see their "Report Three: Comparison of Cutpoints for the 1992 NAEP in Reading Achievement Levels with Those Set by Alternate Means," also summarized in this volume.

level descriptions. The teachers were interviewed about this task, with questions about their reactions to the descriptions, phrases they found helpful in differentiating between levels, and phrases they found difficult to understand or with which they disagreed.

Findings

Results were organized according to the three study questions above.

In their evaluation of the quality of the St. Louis descriptions as a basis for level setting, the investigators found that the Reading Framework did *not* serve as the primary basis of the descriptions, as planned; rather the participants seemed to rely on their experience, intuition, and opinion to construct the descriptions. Dr. Pearson and Dr. DeStefano conclude that extreme inconsistencies (even contradictions) between the St. Louis descriptions and the Reading Framework render them invalid as a basis for level setting. The St. Louis descriptions reflect an out-dated, skills-based, developmental approach to reading. By differentiating between achievement and grade levels on the basis of skills, the St. Louis descriptions are antithetical to the basic underpinnings of the framework, namely that the reading process is common to all readers and that differences are based on variations in text and task.

In regard to the differences between the descriptions used to set achievement levels and those used to report results, the investigators noted that few changes were made in San Diego, mainly, they concluded, because participants felt constrained from doing so. Dr. Pearson and Dr. DeStefano were not privy to the development process of the final version of the descriptions but submitted them for consideration by their expert panel. The experts found that the influence of the Reading Framework was more apparent; for example, the distinction among literary, information, and practical text was incorporated into the descriptions and consistent across levels and grades. However, the experts also found that the descriptions differed in substantive ways, such as in the addition of some descriptions of performance and the deletion of others, substantially altering the content from one version to the other; and in the addition of references to grade-appropriate text and reading situations. The investigators concluded that differences in content and format were significant enough to make it impossible to use the official version of the descriptions to report the results of the 1992 NAEP in reading, which were established on the basis of the achievement levels set in St. Louis.

As to the quality of the official descriptions as a basis for reporting the results of the 1992 NAEP in reading, the investigators find that although these descriptions may seem to represent a marked improvement over those developed in St. Louis, the official descriptions still do not adequately represent current thinking about reading performance. The final descriptions reflect a developmental concept of reading where skills evolve and are distributed differentially across grades and achievement levels, not the notion of reading as an interaction among the reader, the text, and the reading situation. Important factors that might characterize differences in reading achievement, such as text difficulty, type of task, or time demands, are not included in the descriptions. The teachers interviewed about the official descriptions liked them, for the most part, indicating that the descriptions represented high standards for

achievement. They identified several phrases or key words that were confusing, vague, or unfamiliar to them, and many expressed the view that the descriptions were consistent with their personal view of reading but not consistent with most ways of assessing reading.

Summary

Despite improvements in the final version of the achievement levels in reading, sufficient concerns exist to mitigate against their use in future level setting and in subsequent NAEP reporting. In fact, the assumptions that underlie the concept of reading presented in the framework and problems in articulation between the Reading Framework and the NAEP assessment itself call into question the whole idea of using grade-differentiated achievement levels to report performance on the NAEP in reading. The investigators conclude that much more professional deliberation and debate are needed in this area before a long term reporting plan is adopted.

Expert Panel Review of the 1992 NAEP Mathematics Achievement Levels

Edward A. Silver and Patricia Ann Kenney
Learning Research and Development Center
University of Pittsburgh

For the National Academy of Education (NAE), two studies of the review of 1992 NAEP mathematics achievement levels were conducted: one (referred to as Study 4A) was concerned with the establishment of cutpoints by a panel of experts through the use of the achievement levels and mathematics item pool; the other (Study 4B) focused on expert judgments regarding the content and quality of the descriptions of basic, proficient, and advanced performance established by the National Assessment Governing Board (NAGB). Each study is described below, along with the investigators' conclusions and recommendations.

Study 4A Purpose and Procedures

The intent of this study was to bring the judgment of subject-matter experts to bear on the determination of cutpoints associated with each level of performance corresponding to the NAGB achievement-level descriptions for basic, proficient, and advanced achievement. A panel of 14 mathematics content experts was convened to review the items from the 1992 NAEP mathematics assessment and make judgments about those items as they related to the grade-level-specific descriptions of mathematics achievement levels that NAGB provided. The members of the panel

represented a cross section of mathematics content experts, including current and former classroom mathematics teachers, mathematics supervisors, and college- or university-level mathematics educators and mathematicians.

The study was conducted over one and one-half days. After an orientation to review the history of NAEP, its content categories, the achievement-level descriptions, and sample mathematics item types, panelists were given an example of item-scaled stimulus materials and copies of NAEP items. The judgment task presented to this panel of experts differed from the original level-setting procedure in that the panelists had more complete information about the framework and the items and were freer to determine their own procedures. They were told that their task was to compare the items and the achievement-level descriptions in order to classify items with respect to the achievement-level categories, thereby determining cutpoints between the achievement-level categories.

Panelists were separated into six groups: two 4th grade, two 8th grade, and two 12th grade. Each grade-level subgroup was responsible for evaluating approximately half the items in each of the five content strands (i.e., geometry; numbers and operations; measurement; algebra and functions; and data analysis, statistics, and probability) and used copies of NAGB's achievement-level descriptions and each group's grade-level NAEP items. Audiotapes and written records were used to document the sessions. After completing their task, each grade-level group discussed their decisions regarding cutpoints and reached consensus about them for the combined set of NAEP items at the grade level in each content strand.

The panelists also reviewed a subset of multiple-response and extended open-response items, along with the multiple-scale values associated with them. The panelists were asked to comment on any differences or similarities between the cutpoints they had established earlier in the session and the official cutpoints NAGB established as the basis for reporting results. Because the official cutpoints were available only for a composite scale at each grade level rather than for each content strand, the comparison was preceded by discussion of methods for establishing a composite scale based on the group's judgments for each content strand. Finally, panelists also examined the exemplar items chosen by NAGB to accompany the achievement-level descriptions and were asked to comment on the appropriateness of the exemplar items.

Study 4B Purpose and Procedures

Study 4B also involved the 14 members of the expert panel who participated in Study 4A, soliciting their independent judgments about the 1992 NAEP mathematics achievement-level descriptions and the accompanying sets of exemplar items. In particular, panelists were to comment on issues such as whether the achievement levels reflect professionally defensible expectations for student performance in mathematics at each grade level and how well the achievement-level descriptions and exemplar items communicate information about student performance to various constituencies (e.g., educators, the public).

Based on the results of Study 4A and Study 4B and an examination of the entire set of circumstances surrounding the creation of the 1992 NAEP mathematics achievement levels, the investigators cannot recommend without reservation that the descriptions, exemplars, and cutpoints be used to report the test results. In fact, their study pointed to problems related to every aspect of the 1992 NAEP mathematics achievement levels. The results of the study also identified an important source of the problem: the retrofitting of achievement levels to a test that was not designed to be reported with respect to such descriptions. Problems caused by the mismatch between the language of the achievement-level descriptions and the content of the test were encountered in several different ways. The issues for consideration in the design of future efforts to set achievement levels for NAEP are summarized below.

Cutpoints. The cutpoints established through the process used in Study 4A did not result in cutpoints that were congruent with those promulgated by NAGB for reporting the 1992 results. In only one instance did the cutpoint interval determined by the study's expert panel members contain the NAGB cutpoint. The panelists' judgments about basic and advanced level cutpoints differed systematically from those established by NAGB, with the panelists at all grade levels setting lower cutpoints for below basic/basic and higher cutpoints for proficient/advanced. While this systematic pattern of differences suggests a rational basis for the observed differences, on the basis of the data it was impossible to conclude whether the differences between the panelists' judgments and those obtained in St. Louis were the result of variations in procedures used to obtain the judgments, differences in the composition of the groups making the judgments, or some combination of these and other factors.

Panelists were unable to establish cutpoints or cutpoint intervals for all content strands at all grade levels; specifically six below basic/basic cutpoints, one basic/proficient cutpoint, and three proficient/advanced cutpoints were not established. Setting the basic/proficient cutpoints was easy relative to setting either of the other two cutpoints.

Exemplars. The panelists convened in this study judged that many of the exemplars are not strong examples of performance at the indicated levels, finding that most were inappropriate, uninteresting, and not at all exemplary. The panelists also found that some of the examplars for certain achievement levels were not even representative of those levels, according to the performance-scale scores and the NAGB cutpoints.

Descriptions. The panelists in this study felt that the final version of the descriptions (a slightly edited version of the descriptions written in Nantucket) was often difficult to use for the purpose of classifying items in terms of achievement-level descriptions, due to some vagueness of language and some omissions related to topics tested in NAEP, and felt that the classification would have been easier if the St. Louis version had been used. On the other hand, the panelists much preferred the final achievement-level descriptions as statements of achievement aspirations for students.

The Fidelity of Test Content with Achievement-Level Descriptions. The investigators found that, beyond the confusion associated with the process of creating the 1992 mathematics achievement levels, many of the deficiencies and problems identified in this study are due to the mismatch between the content of the NAEP assessment and the aspirations for student performance embodied in the achievement-level descriptions. For realization of the full potential of achievement-level

descriptions as a method of reporting NAEP results, the investigators believe that better articulation between the content of the test and the content of the descriptions must be achieved. This articulation is essential if the test results are to be used to make valid inferences about student performance in mathematics.

Use of Achievement-Level Descriptions. Despite the extensive criticism of the achievement-level descriptions and exemplars for the 1992 NAEP mathematics assessment, the panelists in studies 4A and 4B were generally more favorably inclined toward the *concept* of achievement levels as a method of reporting test results than they were toward the use of scale anchor points. They questioned the adequacy of the scales because of their post-hoc character, their lack of differentiation of performance with grade-level specificity, and their inability to communicate useful general information about expectations for, or observations of, student performance. Although they found the 1992 version of mathematics achievement levels to be seriously flawed, the investigators believe that the achievement-level descriptions, both as statements of achievement aspirations and as categories for reporting student performance, may well be a promising approach in the long term.

When achievement-level descriptions are in place to report results and to guide test construction, they will be able to function as powerful communicators about the mathematics achievement of American students. Most of the problems associated with achievement levels for the 1992 NAEP mathematics assessment are related to the attempt being made to retrofit them to a test that was never designed to have its results reported with respect to such categories. It should not be too late to rectify the situation for future assessments. The investigators suggest that rather than abandoning achievement levels because of the problems associated with their birth, a more planful approach be adopted to their more successful implementation in the future.

Recommendations for the Future. Planning should begin now to put into place achievement-level descriptions that can be used to report the results and merged into the framework used to develop the test items. Careful prespecification of achievement-level descriptions can help provide high fidelity between test content and achievement levels, but the overall quality of the assessment cannot be poor; the implicit model of student proficiency cannot be impoverished.

In order to ensure that the process of setting performance standards for student achievement in mathematics has educational and disciplinary integrity, experts in mathematics content and pedagogy should monitor all substantive aspects of achievement-level creation, framework development, test design, and the reporting of results, in much the same way that the technical aspects are monitored by a cognizant review panel. With an expert panel in place for long term oversight, more broadly representative groups of educators and the general public could be convened at appropriate points to investigate the representativeness of the test, the achievement-level descriptions, and the resulting inferences made about the performance of American students.

Comparison of Teachers' and Researchers' Ratings of Student Performance in Mathematics and Reading with NAEP Measurement of Achievement Levels

Donald H. McLaughlin, Phyllis A. DuBois, Marian S. Eaton, Dey E. Ehrlich, Fran B. Stancavage, Catherine A. O'Donnell, and Jin-ying Yu
American Institutes for Research

Lizanne DeStefano, David Pearson, Diane Bottomley, Cheryl Ann Bullock, Matthew Hanson, and Cindi Rucinski
University of Illinois at Urbana-Champaign

The process of setting achievement levels for the 1992 National Assessment of Educational Progress (NAEP) resulted in two primary products: (1) verbal descriptions, or characterizations, of performance at three different achievement levels—basic, proficient, and advanced—separately for 4th, 8th, and 12th grade, and (2) NAEP scale-score cutpoints between the levels. The presumption was that the school performance of students who obtained scores on NAEP between the lower and upper boundaries of a level would be likely to be described by their teachers as matching the verbal description for that level.

Because future results of NAEP assessments will be reported in terms of the percentages of the nation's students who show evidence of performing at the three achievement levels, it is important to know whether the various audiences for those reports will interpret the words *basic, proficient,* and *advanced* in the same way.

Purpose of the Study

This study focuses on the effectiveness of the achievement-level descriptions and asks this question:

> Would a group of teachers similar to those who participated in setting the 1992 NAEP achievement levels, using the verbal descriptions of the levels, categorize their own students' classroom performance as advanced, proficient, basic, and below basic in such a way that they match the categories of NAEP scale scores assigned to those students' test performance?

The achievement-level cutpoints had been determined by the National Assessment Governing Board (NAGB) through a procedure that focuses on individual item performance. This study asks whether the validity of those levels extends to holistic judgments of performance of students in school, as seen by their teachers. Do the verbal descriptions convey sufficient information that readers of NAEP reports will accurately interpret what is meant by statements like "fewer than 20 percent of the nation's eighth-grade students are performing proficiently in mathematics?"

To assess the extent to which teachers' classifications of their students' school performance as advanced, proficient, or basic is colored by the range of students in their particular school, an independent assessment of a sample of the students' performance was obtained. For this purpose, performance validation protocols were developed for one-on-one interviews with students who were assessed and rated by trained researchers. To gather NAEP scores, the study was "piggy-backed" on a previously planned NAEP data-collection effort (the 1993 field test of items to be used in the 1994 assessment), and an equating sample of 1992 forms was included within the field-test classrooms. The cooperation of the NAEP contractors, Educational Testing Service, Westat, and National Computing Systems, contributed significantly to the timely completion of the study.

Schools were recruited for participation after they had agreed to participate in the 1993 NAEP field test. They were clearly informed that participation in this study was voluntary, that they could refuse without jeopardizing their participation in NAEP, and that their school would receive $100 to use for instructional materials if they participated. Over 90 percent of the contacted schools agreed to participate.

In the interests of economy, the study was carried out only for fourth- and eighth-grade reading and mathematics. Subjects were 5,035 students in 171 schools, and teachers' ratings were obtained on 4,798 of these students. Due to absences and other reasons, such as failure to attempt any items, NAEP scores were obtained for only 4,063 of the students whom teachers rated, including 489 scores on 1992 assessment booklets. For fourth grade, the final database included 1,001 students in 48 schools for mathematics and 913 students in 39 schools for reading; for eighth grade, 1,145 students in 38 schools for mathematics and 1,004 in 46 schools for reading.

Primary materials were the NAEP item booklets (several 1993 booklets, plus two 1992 booklets for each grade and subject area) and NAEP Administration Schedules (a form which listed each participant's name and booklet identification and included a column for teachers to use in indicating their ratings of student performance). Teachers were given a set of instructions for entering the ratings and a copy of the official versions of the three NAEP achievement levels for their grade and subject area.

To validate the teachers' ratings through 20-minute on-site one-on-one interviews with participating students, detailed protocols were developed. For mathematics, they consisted of five or six problems, adapted by staff from problems presented in the framework of the National Council of Teachers of Mathematics (NCTM). Protocols were organized to present each problem first as a whole, to allow the student to work it out without scaffolding, then, as needed, with a series of probes to determine what a student who could not solve the entire problem could do on parts of it. For reading, each student selected a text to read silently for 10 minutes; then a researcher engaged the student in conversation about it, using questions and probes that focused on the meaning the text conveyed to the student, the student's evaluation of the text, the student's reactions to the characters, and the student's opinions about the author's work.

Sixteen sites were selected for on-site visits to validate teachers' ratings for each grade level and subject area. The sites were selected to cover all regions of the nation and different types of communities; otherwise, for travel convenience. At each visited site, eight students were selected for interviews, a boy and girl at each of the four teacher-rating levels when possible. Ratings from both teachers and researchers were obtained for 473 students.

Analysis and Findings

Four separate analyses were undertaken: one each for fourth-grade and eighth-grade reading and mathematics. The main analysis was based on the construction of a four-by-four table of classifications according to teachers and assessment results, using the official NAEP achievement-level boundary scores. A secondary analysis was made of the extent to which there is coherence between the teacher judgments and the NAEP assessment of the achievement levels.

Cutpoints were identified as the points on the scale at which misclassification rates between categories (e.g., between proficient and not proficient) were equal in both directions. Cutpoints were estimated separately for each school, and averages were computed across the approximately 40 participating schools in each grade and subject.

The major result is that substantially more students were considered by teachers to have demonstrated each level of performance in the classroom than would have been estimated using the NAEP scores and achievement-level cutpoints. For example, teachers considered 39 percent of the participating fourth graders to be proficient (or better) in reading, compared to a NAEP estimate of 11 percent. This could be translated into a cutpoint difference of 29 points (243 minus 214). A similar result was obtained for each of the 12 comparisons between teachers' ratings and NAEP scores. Each difference was statistically significant, with lower cutpoints set by participating teachers than by NAEP. Furthermore, researchers' ratings in reading, and in one grade for mathematics, agreed with teachers' ratings. The exception to the general result was in eighth-grade mathematics, in which researchers' ratings more closely agreed with NAEP results than with teachers' ratings.

To address the extent to which teachers and researchers were making ratings in terms of the same underlying performance scales as measured by NAEP, the intercorrelations of classifications were computed. All correlations of ratings with NAEP scores were significantly positive; however, the patterns differed somewhat between reading and mathematics. In reading, researchers' ratings were more highly correlated with teachers' ratings than either were with NAEP scores. In mathematics, on the other hand, researchers' ratings were the least correlated of the three measures. Generally, there appeared to be *no* tendency of the teachers in this study to under-rate performance in schools with many high-performing students and over-rate performance in schools with few high-performing students.

Comparisons of Student Performance on NAEP and Other Standardized Tests

Elizabeth Hartka
American Institutes for Research

The purpose of this study was to determine the concurrence between the National Assessment of Educational Progress (NAEP) achievement-level estimates of "advanced" 12th graders and other independent estimates of "advanced" 12th graders. The motivation for the investigation was partially provided by the "generic" definition of advanced performance on the NAEP proposed by the National Assessment Governing Board (NAGB):

> For 12th grade the advanced level shows a readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement. As data become available, it...may be related to Advanced Placement and other college placement exams.[1]

It seems reasonable to assume that proportions of students performing at the "advanced" level on the 12th grade NAEP should be comparable to proportions of "advanced" students determined by performance on tests for college-bound students, such as the Advanced Placement (AP) or the Scholastic Aptitude Test (SAT). For this reason, the investigator examined proportions of 12th-grade students who performed at high levels on the Verbal and Mathematics subscales of the SAT and the College Board's Advanced Placement tests in Calculus (forms AB and BC), English Literature, and English Language, and compared the proportions of students performing at high levels to the percentage of 12th graders who performed at the advanced level on the NAEP mathematics or reading assessments.

Procedures

Frequency distributions for the SAT and AP tests were examined, and the numbers scoring above various levels judged to be "advanced" were computed. Then, the number of students judged to be advanced was divided by the number of high school graduates in the appropriate year to estimate the percentage of advanced students. The number of high school graduates for school years ending in 1990 and 1991 were calculated using data obtained from the National Center for Education Statistics; for the school year ending in 1992, an estimate was obtained from census data.

[1] National Assessment Governing Board, *Achievement Level Options for the NAEP Mathematics Assessment: 1990 Trial* (Washington, D.C.: Author, May 10, 1991).

Scholastic Aptitude Test data. Frequency distributions of student scores on the SAT were obtained from the Educational Testing Service reports.[2] The 1992 report described the examination this way:

> The SAT is a 2.5 hour, multiple-choice test that measures developed verbal and mathematical reasoning abilities related to successful performance in college. SAT scores are intended to be used with the secondary school record and other information about the student in assessing readiness for college-level work.[3]

Student performance on the SAT is reported on a scale from 200 to 800, with a standard error measurement of approximately 30 points. A score of 600 on the SAT is traditionally considered to be high; percentages of students scoring above 600 are reported annually. The means for the Verbal section of the SAT are about 50 points lower than the means for the corresponding Mathematics section in any given year. To make scale scores for the two scales more equivalent, the College Board recently made a decision to rescale the Verbal test. Given the discrepancy between Mathematics and Verbal mean scores, the investigator chose a cutoff of 550 to determine advanced performance for the Verbal SAT scores, with 600 for Mathematics SAT scores.

Advanced Placement data. Summary reports of AP performance were used to determine the numbers of students scoring at high levels. AP examinations are given in May each year at participating schools and allow students performing at certain levels to earn college credit or advanced placement. Usually students who take the AP exams have participated in an AP course at their high schools, with course descriptions, curricular outlines, teaching guides, and examinations all prepared by the AP Program. The examinations are graded by panels of experts, using a 1-5 score, ranging from 1, no recommendation, to 5, extremely well qualified.

In the judgment of the investigator, students scoring above 3 ("qualified") were probably overqualified as "advanced" by NAGB's criterion of showing "readiness for rigorous college courses," since they had already (successfully) completed college-level work. In 1992, 378,692 candidates took 566,036 AP Examinations. These candidates represented 9,730 U.S. secondary schools. Total program participation also included 14,107 examinations taken by 9,450 candidates from 461 schools abroad.[4]

Findings

The SAT scores for three consecutive school years indicate that on the Verbal section, 5.7 percent of high school graduates scored at 550 or higher in 1990 and 5.8 percent in 1991 and 1992. On the Mathematics section, 7.1 percent of high school graduates

[2] Educational Testing Service, *College Bound Seniors: 1990 Profile of SAT and Achievement Test Takers; 1991 Profile of SAT and Achievement Test Takers;* and *College Bound Seniors: 1991 Profile of SAT and Achievement Test Takers* (Princeton, NJ: Author).

[3] Ibid., inside front cover.

[4] Advanced Placement Program of the College Board, *1992 Advanced Placement Program National and Alaska Summary Report* (New York: Author, 1992), inside front cover.

scored 600 or higher in 1990, 7.0 percent in 1991, and 7.5 percent in 1992. These percentages differ from proportions of SAT test takers scoring above these cutpoints. Percentages of test takers would be much higher, since the SAT takers are a much more select population than the population of high school graduates in the United States. The estimate are crude and almost certainly underestimate the percentage of 12th-grade students capable of scoring at a given level because many college-bound high school students take the American College Test (ACT) and not the SAT.

The SAT data suggest that NAEP advanced cutpoints may have been set very high. In mathematics in 1992, only 2 percent of U.S. 12th graders scored at the advanced level on NAEP, whereas at least 7.5 percent of high school graduates scored at 600 or better on the SAT. On the SAT Verbal test, which measures vocabulary, verbal reasoning, and reading comprehension, 5.8 percent scored at a high level (550) in comparison to the finding from NAEP that only 3.2 percent of 12th graders are advanced.

The counts of AP test takers included students from all grades, both public and private high schools and colleges, who took the examinations in May 1992. For mathematics, the Calculus AB and BC tests were used. The Calculus AB test covers introductory differential and integral calculus; Calculus BC includes advanced topics in integral calculus and sequences and series. Candidates were permitted to take only one of the Calculus examinations. For English, the English Language and Composition test and English Literature and Composition tests were used; each is intended to cover a full-year introductory college English course. Candidates could register for both English examinations, but the amount of overlap for 1992 was very small (only 167 students took both exams).

In mathematics, an estimated 2.5 percent of high school graduates earned a score of 3 or greater on the AP Examination (which would qualify them for college credit) in 1992; in English, it is 3.8 percent. Like the SAT results, these percentages should be treated as underestimates, the investigator cautions, since not all potentially advanced students in the United States have access to AP programs or tests. AP Examinations in any subject area are available in only 46 percent of U.S. secondary schools.

Taken together, the approximate comparisons provided by the SAT and AP tests suggest that the NAEP advanced cutpoints in reading and mathematics may have been unreasonably high. From the perspective of these other tests, a higher percentage of 12th graders would be identified as advanced.

Comparing the NAEP Trial State Assessment Results with the IAEP International Results

Albert E. Beaton and Eugenio J. Gonzalez
Boston College

Policymakers and the general public want to know how American students compare to the students of other countries today, whether the national education goal of making U.S. students first in the world in science and mathematics achievement by the

year 2000 is being approached, and, ultimately whether it will be attained. Information about the past has come from the International Assessment of Educational Progress (IAEP), which collected data and compared the achievement of American students in science and mathematics to that of students from other nations. Information in the future will come from the Third International Mathematics and Science Study (TIMSS), which will conduct international surveys of both mathematics and science in 1995 and 1998.

With the introduction by the National Assessment Governing Board (NAGB) of achievement levels for student performance on NAEP, it is of interest to know how well students in other countries would meet these standards. No data are available at present to address this question directly. This study attempts to provide approximate answers to the questions of how American students compare to foreign students in mathematics and how well foreign students can meet the NAGB mathematics standards, by using available data, in particular the 1990 NAEP data (including the Trial State Assessment [TSA] data), and the IAEP 1991 data. The questions about achievement in mathematics and science of American students cannot be answered rigorously without full-scale studies such as the TIMSS survey; the approximate answers for mathematics produced by this study require many assumptions which are made explicit. Under these assumptions, the comparisons of state to international performance are made.

The NAEP and IAEP Data

Two data sets come reasonably close to being able to answer the study questions, and the investigators use them to make inferences: NAEP and IAEP. Both samples measure the mathematics achievement of students at the middle school level. In 1990 NAEP collected data on *national* samples of students for reporting their proficiency in reading, writing, mathematics, and science. Data were collected for students who were either in the 4th, 8th, or 12th grades or were 9, 13, or 17 years old. The NAEP TSA collected and published distributions of the mathematical achievement of eighth-grade public school students in individual *states*. Taking part were 37 states, the District of Columbia, and two territories. The NAEP database also included an eighth-grade, winter, public school sample which is a NAEP subsample that is directly comparable to the state samples.

IAEP was conducted in 1991 by the Educational Testing Service under contract from the National Center for Education Statistics. The United States, 19 other countries, 9 Canadian provinces, and the State of Colorado participated in the survey (with Colorado participating as if a nation). The sample of students included 9- and 13-year-olds who were assessed in mathematics and science. The United States and eight other countries also assessed geography.

Data collected in the TSA in 1990 was for eighth-grade mathematics only; no science data are available for individual states. Because American eighth graders are 13 years old, for the most part, the overlap with the population samples internationally is considerable. Other major similarities and differences between the two surveys are the following: (1) The TSA took place in February 1990; the IAEP in March 1991. (2) Both the TSA and the IAEP used the same assessment framework in specifying the contents of their mathematics test; the tests should therefore be similar in content. (3) The 1990

TSA and IAEP mathematics assessment had no items in common. (4) The IAEP required an hour and 15 minutes; the TSA, only 45 minutes. (5) Both the TSA and IAEP scaled their data using similar, but not identical, techniques. (6) The TSA sampled at grade 8; the IAEP, at age 13. (7) The TSA sampled public school students only; the IAEP sampled both public and private schools. (8) Both the TSA and IAEP data sets are well documented and available as public-use data files.

These differences make it clear that the assumptions for formal equating and rigorous comparisons are not met. The investigators instead did a statistical moderation of the two samples. To give the results meaning, they also made the following assumptions:

- The two assessment instruments measure the same mathematical skills.
- The tests measure the mathematical skills with equal accuracy.
- The relationship between the performance of public school eighth graders in a particular state to the public school eighth graders in the United States as a whole is the same as the relationship of all 13-year-olds in that state to all 13-year-olds in the United States as a whole.
- The relationship between eighth graders and 13-year-olds specified above is the same in other countries as it is in the United States.
- No major changes in relative student achievement occurred between 1990 and 1991.

Procedure

To carry out the statistical moderation between the TSA and IAEP data, the investigators made a simple transformation of the IAEP scale into the NAEP mathematics composite scale used by the TSA. The choice was made to transform the IAEP scale into the NAEP scale because the NAEP scale is better known and the achievement levels are defined on that scale. Three sources of data were used in the analysis: NAEP, TSA, and IAEP.

The procedure had four steps. (1) Linear equating methodology was used to develop an equation for transforming the IAEP mathematics scores into the NAEP composite mathematics scale, using IAEP United States sample and the NAEP eighth-grade, winter, public school national sample. (2) IAEP plausible values for all countries were transformed to the NAEP metric. (3) The average mathematics achievement score and the percentage above and below the NAGB achievement levels was computed for each country and for each state. (4) State-versus-country comparisons were prepared.

Results

The results of this study reflect previous, widely published findings regarding the relatively low position of U.S. students in mathematics achievement in comparison to the achievement of students in other countries. Because of the relatively low U.S. average, most state averages would be expected to be lower than those of other countries. However, some individual states had higher averages than many other countries; for example, students in North Dakota and Montana outperformed students in most of the IAEP countries.

For student performance at the advanced level, perhaps the most striking observation is the large percentage in Taiwan (24.4 percent) and Korea (15.6 percent), when compared to the percentage in the United States (2.9 percent). Even the states with the highest estimates (Montana and Virginia, both with 4.1 percent) are nowhere near as high. The largest percentages for European countries are 7.6 percent for Hungary and 5.9 percent for the former Soviet Union.

At the proficient or advanced level, the percentage of students is also telling. Over half of the students in Taiwan (54.1 percent) and Korea (52.2 percent) have reached these levels, as have over 40 percent of the students in the former Soviet Union, Switzerland (41.8 percent), and Hungary (40.1 percent), compared to only 17.3 percent of the students in the United States. The states with the largest percentages reaching these levels are North Dakota (33.8 percent) and Montana (32.6 percent).

Forty-two percent of U.S. students score at the below basic level, a number far greater than in the top-scoring countries, many of which (Taiwan, Korea, the former Soviet Union, Switzerland, and Hungary) had fewer than 20 percent in this category. Of the states, only North Dakota and Montana had fewer than 20 percent of their students below the basic level. The data clearly indicate that students in other industrialized countries are outperforming U.S. students in mathematics at the 13-year-old level.

Methodological Discussion

To determine how reasonable their results were, the investigators carried out a check, using the data from the state of Colorado, which elected to be in both the TSA and in the IAEP surveys. If all of the investigators' assumptions were correct, then the Colorado results from the IAEP study should be the same as the results from the TSA study, except for sampling and measurement error. The TSA distribution was slightly higher on the NAEP scale than the IAEP-transformed distribution—a difference the investigators considered reasonably small but one that is statistically significant. The difference was large enough to alter somewhat the rank-order of the various states and countries, but a major difference, that would reorder states that were already significantly different, was not found. The differences may be larger for other countries. The investigators' judgment is that these approximations are serviceable for some purposes until better data are available.

An Evaluation of the 1992 NAEP Reading Achievement Levels, Report Three: Comparison of Cutpoints for the 1992 NAEP Reading Achievement Levels with Those Set by Alternate Means

David Pearson and Lizanne DeStefano
University of Illinois at Urbana-Champaign

This investigation compared the cutpoints defining the achievement levels of basic, proficient, and advanced for the 1992 National Assessment of Educational Progress (NAEP) in reading, which were set in St. Louis by using a modified Angoff procedure, with the cutpoints set by using an alternative procedure. The primary purpose was to establish the degree of concurrence as an indicator of the validity of the NAEP achievement levels. A secondary benefit was the opportunity to learn more about the impact of participant characteristics and the nature of information available on the process of level setting.

Method

A 1-day alternative level-setting session was held on the campus of the University of Illinois, Urbana-Champaign, with the investigators serving as facilitators for 17 participants. The level-setting process used at this session differed in at least five ways from the modified Angoff approach used to set the NAEP achievement levels: composition of the expert panels, use of pre-established achievement-level descriptions, information available to participants, nature of the consensus process, and role of the content-area experts and facilitators. Each is summarized below.

Composition of the expert panels differed in that all participants in Champaign were educators with extensive experience in teaching or supervising reading at specific grade levels; all were familiar with state and national assessment initiatives, including NAEP; and all were active members of at least one professional organization in reading or language (e.g., International Reading Association, National Reading Council, National Council of Teachers of English). In contrast, the group that set the NAEP achievement levels included representatives of the general public, teachers, and nonteacher educators. The investigators, who observed the level-setting session in St. Louis, had found that few of the participants were familiar with the Reading Framework prior to the meeting, and the diverse experiences and knowledge of these participants made consensus difficult to achieve. The result was a reliance on experience, intuition, and opinion, so a homogeneous group was selected for the Champaign session.

Participants in St. Louis used ***pre-established achievement-level descriptions;*** participants in Champaign were ask to use the NAGB descriptors to define the achievement levels and to serve as the basis of identifying cutpoints. The final version of the descriptors, developed after the St. Louis meeting, was used, since it was considered the most likely to be used in reporting.

Information available to participants in Champaign was very different from that provided in St. Louis. The St. Louis rating process was iterative, conducted in three rounds. During the first round, participants at each grade level were given no information other than the reading passages, items, rubrics, and the descriptions for that grade level. They were asked to estimate the percentage of students at the borderline of each achievement level who would get the item right. In the second round, participants were able to adjust their first round ratings after receiving information on item difficulty in the form of p-values and information on interjudge consistency. During the third round, participants were given access to intrajudge consistency indicators, in addition to interjudge consistency information. In Champaign, participants were also given items, passages, and rubrics; however, instead of p-values they were given a set of item maps that provided a visual representation of the relative difficulty of all items of each grade level arrayed on the NAEP performance scale (0-500). Multiple-choice, short-answer, and extended-response items were also represented on the map.

The nature of the consensus process changed from limited consultation to group decisionmaking. In St. Louis, consensus was achieved by aggregating and averaging the individual ratings of all participants. They worked alone during Rounds 1 and 2 and were allowed to consult with tablemates only in Round 3. In Champaign, the process was iterative, with a goal of obtaining consensus across grade-level groups of participants through continual negotiation and discussion. Participants were assigned to one of the three grade levels (4th, 8th, and 12th), and two teams of two or three members were formed at each grade level. Each team was given four item maps for its grade level; collectively the maps represent all of the items at the grade level. Participants were asked to examine a single item map, consider the descriptors, the reading passages, and the items, and to identify the point at which item difficulty and content was indicative of borderline performance at each achievement level. For the first map, they identified three cutpoints (below basic/basic, basic/proficient, and proficient/advanced). Participants were then instructed to turn to the second map, align it with the first, and use the additional information provided by the new items to make adjustments in their original cutpoints. The third and fourth maps were used to further refine the cutpoints, until each team had used information from all items to inform the identification of the cutpoints. After both teams at a grade level had arrived at a final set of cutpoints, they came together as a grade-level group, presented their cutpoints and attempted, through discussion , to reach consensus on a single set of cutpoints that all members of the grade-level group could endorse.

The role of the content-area experts and facilitators changed from minimum interaction in St. Louis, to close involvement in Champaign. The St. Louis meeting kept interactions between content-area experts and participants to a minimum to allow participants maximum autonomy when developing the descriptions and assigning ratings to items. Content-experts revised interim and final versions of the descriptions and distributed them to participants; two of the three left before the rating process began. The Champaign content-area experts were highly involved in all aspects of the rating process: they facilitated discussions around interpretations of the descriptions, suggested strategies for resolving conflicts, assisted in interpreting item maps, and kept each grade on schedule.

Comparisons were made of the grade-level cutpoints determined by both level-setting sessions (St. Louis and Champaign) and the range of cutpoints established by the two teams at each grade level before coming together and negotiating a grade-level cutpoint. When compared with cutpoints developed in St. Louis, the Champaign cutpoints were generally higher, except in the case of 12th-grade basic and proficient. When analyzed for cross-level and cross-grade consistency, the Champaign cutpoints show steady upward trends as grade level and achievement increase, again with the exception of 12th grade basic and proficient, which were set quite near their 8th grade counterparts.

In reviewing the impact of characteristics of participants and process on level setting, the investigators concluded that the use of a heterogeneous group of participants in Champaign enabled the group to comprehend the process and establish strategies for defining cutpoints much more efficiently than their St. Louis counterparts. The development of descriptors was not demonstrated to be a necessary component of level setting, and the investigators believe that this finding lends support to the idea of a level-setting process in which the descriptors are developed by content experts in conjunction with the content framework and then applied to the level-setting process. Regarding the information available to participants, it was clear to the investigators that the nature of information available in Champaign inextricably linked the level-setting process to the content of the 1992 NAEP in reading, and they voiced concerns when items did not address aspects of the descriptors (e.g., critical evaluation or analysis of author's intent). In contrast, St. Louis participants rated items individually and raised few, if any, concerns about the representativeness of the items. Consensus was elusive in both Champaign and St. Louis; even with the carefully chosen and fairly homogeneous group in Champaign, consensus was not achieved at the 12th-grade level. The involvement of content specialists in Champaign was sought out and viewed positively by participants.

It may not be surprising, the investigators conclude, that two different level-setting processes, relying on different information and conducted with different participants, resulted in different cutpoints to define achievement levels associated with the 1992 NAEP in reading. Because so many factors changed from one session to the next, the source or sources of the variation are impossible to isolate, and it cannot be argued that the levels established by the modified Angoff method are more or less valid than those set in Champaign. The important finding is that achievement levels are not objective, highly visible, commonly construed constructs. They are constructed through and dependent upon the interaction among participants, information about the test and student performance, and the level-setting process. Variations in any of these factors can result in quite different representations of achievement.

An important indicator of the validity of achievement levels, then, is the credibility of the process used to set them. In their first report, the investigators concluded that the modified Angoff process was sufficiently flawed to render the resulting achievement levels invalid. In the alternate process, homogeneous raters, the use of pre-established achievement-level descriptors, reductions in the amount and complexity of

information available to participants, and increased involvement of the content specialists seemed to offer improvements over the modified Angoff. Consensus was not achieved universally, however, a fact that still brings into question the notion of using these or any other achievement levels to report NAEP data.

The investigators' evaluation of both the St. Louis and Champaign level-setting sessions leads them to conclude that while the development of standards or achievement-level descriptions that are highly valued by the field and closely linked to the content of the assessment is an absolutely critical starting point in this approach to NAEP reporting, we have not yet reached that point.

Notes

Alexander, Lamar and James, H. Thomas. *The Nation's Report Card: Improving the Assessment of Student Achievement.* Washington, D.C.: The National Academy of Education, 1987.

American College Testing Program. *Descriptions of Mathematics Achievement Levels-Setting Process and Proposed Achievement Level Definitions.* Iowa City, IA: Author, April 13, 1992.

Beaton, A.E. and Gonzalez, E.J. "Comparing the NAEP Trial State Assessment Results with the IAEP International Results," in *Setting Performance Standards for Student Achievement: Background Studies.* Stanford, CA: The National Academy of Education, 1993.

Bourque, M.L. "The NAEP Achievement Level Setting Process for the 1992 Mathematics Assessment," *Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics.* E.G. Johnson, J. Mazzeo, and D.L. Kline. Washington, D.C.: National Center for Education Statistics, 1993.

Bourque, M.L. and Garrison, H. *The LEVELS of Mathematics Achievement: Initial Performance Standards for the 1990 NAEP Mathematics Assessment.* Washington, D.C.: National Assessment Governing Board, 1991.

Bruce, B., Osborn, J., and Commeyras, M. "The Content and Curricular Validity of the 1992 National Assessment of Educational Progress in Reading." Stanford, CA: The National Academy of Education, forthcoming.

Council of Chief State School Officers. "Comments of the Council of Chief State School Officers on 'Setting Goals for the National Assessment.'" Presentation at a National Assessment Governing Board Forum, January 25, 1990.

Finn, Jr., C.E. News release. Washington, D.C.: National Assessment Governing Board, December 12, 1989.

Ford, Honorable William D. and Kildee, Honorable Dale E.: Personal communication to Eleanor Chelimsky at U.S. General Accounting Office. Washington, D.C., October 7, 1991.

Glode, Michael: Personal communication to Robert Glaser and Robert Linn. Washington, D.C., February 4, 1992.

Goals 3/4 Technical Advisory Group meeting, Council of Chief State School Officers. Washington, D.C., April 29, 1993.

Hartka, E. "Comparisons of Student Performance on NAEP and Other Standardized Tests," in *Setting Performance Standards for Student Achievement: Background Studies.* Stanford, CA: The National Academy of Education, 1993.

Jaeger, R.M. "Certification of Student Competence," *Educational Measurement*. 3rd ed., Ed. R.L. Linn. New York, NY: American Council on Education and Macmillan, 1989.

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

Linn, R.L., Koretz, D.M., Baker, E.L., and Burstein, L. *The Validity and Credibility of the Achievement Levels for the 1990 National Assessment of Educational Progress in Mathematics*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, January 1991.

Linn, R.L., Shepard, L., and Hartka, E. "The Relative Standing of States in the 1990 Trial State Assessment: The Influence of Choice of Content, Statistics, and Subpopulation Breakdowns," in *Studies for the Evaluation of the National Assessment of Educational Progress (NAEP) Trial State Assessment*. Stanford, CA: The National Academy of Education, 1992.

Livingston, Samuel A.: Personal communication with author, April 29, 1993.

McLaughlin, D.H. "Validity of the 1992 NAEP Achievement-Level-Setting Process," in *Setting Performance Standards for Student Achievement: Background Studies*. Stanford, CA: The National Academy of Education, 1993.

McLaughlin, D.H. "Order of Angoff Ratings in Multiple Simultaneous Standards," in *Setting Performance Standards for Student Achievement: Background Studies*. Stanford, CA: The National Academy of Education, 1993.

McLaughlin, D.H. "Rated Achievement Levels of Completed NAEP Mathematics Booklets," in *Setting Performance Standards for Student Achievement: Background Studies*. Stanford, CA: The National Academy of Education, 1993.

McLaughlin, D.H., DuBois, P., Eaton, M., Ehrlich, D., Stancavage, F.B., O'Donnell, C., Yu, J., and DeStefano, L. "Comparison of Teachers' and Researchers' Ratings of Students' Performance in Mathematics and Reading with NAEP Measurement of Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies*. Stanford, CA: The National Academy of Education, 1993.

Messick, S., Beaton, A., and Lord, F. *National Assessment of Educational Progress Reconsidered: A New Design for a New Era*. NAEP Report No. 83-1. Princeton, NJ: Educational Testing Service, 1983.

Mullis, I.V.S., Dossey, J.A., Owen, E.H., and Phillips, G.W. *NAEP 1992 Mathematics Report Card for the Nation and the States*. Washington, D.C.: National Center for Education Statistics, 1993.

National Academy of Education. *Assessing Student Achievement in the States*. Stanford, CA: Author, 1992.

National Assessment Governing Board. "Discussion Paper on the Future of the National Assessment of Educational Progress." Washington, D.C.: August 1992.

National Center for Education Statistics, I.V.S. Mullis et. al. *The STATE of Mathematics Achievement.* Washington, D.C.: Author, June 1991.

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

National Council on Education Standards and Testing. *Raising Standards for American Education: A Report to Congress, the Secretary of Education, the National Education Goals Panel, and the American People.* Washington, D.C.: U.S. Government Printing Office, January 24, 1992, ISBN 0-16-036087-8.

National Education Goals Panel. *Measuring Progress Toward the National Education Goals: Potential Indicators and Measurement Strategies.* Washington, D.C.: Author, March 25, 1991.

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

Pashley, P.J. and Phillips, G.W. *Toward World-Class Standards: A Research Study Linking International and National Assessments.* Princeton, NJ: Educational Testing Service, June 1993.

Pearson, D. and DeStefano, L. "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report One: A Commentary on the Process," in *Setting Performance Standards for Student Achievement: Background Studies.* Stanford, CA: The National Academy of Education, 1993.

Pearson, D. and DeStefano, L. "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report Two: An Analysis of the Achievement-Level Descriptions," in *Setting Performance Standards for Student Achievement: Background Studies.* Stanford, CA: The National Academy of Education, 1993.

Pearson, D. and DeStefano, L. "An Evaluation of the 1992 NAEP Reading Achievement Levels, Report Three: Comparison of Cutpoints for the 1992 NAEP Reading Achievement Levels with Those Set by Alternate Means," in *Setting Performance Standards for Student Achievement: Background Studies.* Stanford, CA: The National Academy of Education, 1993.

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

Public Law 100-297, Part C, Section 3403 6A: 1988.

Public Law 95-561, Section 1242: November 1, 1978.

Rothman, R. "NAEP Plan to Set Performance Goals Questioned." *Education Week* 9(19). January 31, 1990.

Rothman, R. "NAEP to Create Three Standards for Performance." *Education Week* 9(35). May 23, 1990.

Silver, E.A. and Kenney, P.A. "Expert Panel Review of the 1992 NAEP Mathematics Achievement Levels," in *Setting Performance Standards for Student Achievement: Background Studies*. Stanford, CA: The National Academy of Education, 1993.

Silver, E.A. and Kenney, P.A. "The Content and Curricular Validity of the 1992 NAEP TSA in Mathematics." Stanford, CA: The National Academy of Education, forthcoming.

Stufflebeam, D.L., Jaeger, R.M., and Scriven, M. *Summative Evaluation of the National Assessment Governing Board's Inaugural 1990-91 Effort to Set Achievement Levels on the National Assessment of Educational Progress*. Washington, D.C.: National Assessment Governing Board, August 1991.

Technical Review Panel: Personal communication to Gary Phillips. Washington, D.C., March 1993.

Truby, Roy. "The Future of NAEP as the Nation's Report Card." Washington, D.C.: National Assessment Governing Board, November 29, 1989.

Truby, Roy. "Staff Paper on Setting Goals for the National Assessment." Washington, D.C.: National Assessment Governing Board, December 8, 1989.

Truby, Roy. "Setting Appropriate Achievement Levels for the National Assessment of Educational Progress." Washington, D.C.: National Assessment Governing Board, May 10, 1990.

U.S. Department of Education. *A Nation at Risk: The Imperative for Educational Reform*. Washington, D.C.: Author, 1983.

U.S. General Accounting Office. *National Assessment Technical Quality*. Report No. GAO/PEMD-92-22R. Washington, D.C.: Author, March 1992.

U.S. General Accounting Office. *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations*. Report No. GAO/PEMD-93-12 Washington, D.C.: Author, June 1993.

Acronyms Appearing in this Report

ACT	American College Testing
AP	Advanced Placement
CCSSO	Council of Chief State School Officers
EIAC	Education Information Advising Committee
ETS	Educational Testing Service
GAO	General Accounting Office
IAEP	International Assessment of Educational Progress
IRT	Item Response Theory
LRDC	Learning Research and Development Center
KIRIS	Kentucky Instructional Results Information System
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 Education Standards and Testing
NCRMSE	National Council for Research in Mathematics and Science Education
NCS	National Computer Systems
NCTM	National Council of Teachers of Mathematics
NEGP	National Education Goals Panel
SAT	Scholastic Aptitude Test
TIMSS	Third International Mathematics and Science Study
TRP	Technical Review Panel
TSA	Trial State Assessment