DOCUMENT RESUME

ED 338 654 TM 017 419

TITLE	Setting Appropriate Achievement Levels for the National Assessment of Educational Progress: Policy Framework and Technical Procedures.
INSTITUTION	National Assessment Governing Board, Washington, DC.
PUB DATE	10 May 90
NOTE	38p.
PUB TYPE	Guides - Non-Classroom Use (055) -- Reports - Evaluative/Feasibility (142) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Academic Achievement; *Academic Standards; Achievement Tests; *Educational Assessment; Educational Policy; Grade 4; Grade 8; Grade 12; Intermediate Grades; Knowledge Level; Longitudinal Studies; Measurement Techniques; *National Programs; National Surveys; *Policy Formation; Secondary Education; Student Evaluation; Testing Programs; Test Use
IDENTIFIERS	Angoff Methods; *National Assessment of Educational Progress; *Standard Setting

ABSTRACT

The National Assessment Governing Board (NAGB) is responsible for improving the form and use of the National Assessment of Educational Progress (NAEP). The NAGB is beginning to define achievement levels to state clearly what students should know and be able to do at key grades in school. This report creates a policy framework, definitions, and technical procedures for establishing these achievement levels. The report is divided into three sections: policy framework, technical procedures, and display of NAEP results in terms of achievement levels. The following three levels are to be established for each grade and subject tested: (1) proficient, a solid academic performance for grades 4, 8, and 12; (2) advanced, signifying superior performance beyond mastery at grades 4, 8, and 12; and (3) basic, a demonstration of partial mastery of knowledge and skills that are fundamental for proficient work at grades 4, 8, and 12. The NAGB intends to use this framework for reporting results for newly developed assessments for 1992 and subsequent years. An ad hoc advisory panel is to be appointed to assist in defining the levels, drawing on a number of assessments and studies. The second part of this report, technical procedures to be used, includes a modified Angoff procedure for standard setting. Appendices to the second section provide sample forms for use in the process. The third section of this document contains four sample graphics as potential ways of reporting achievement level information. (SLD)

National Assessment Governing Board

National Assessment of Educational Progress

SETTING APPROPRIATE ACHIEVEMENT LEVELS

For The

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Policy Framework and Technical Procedures

APPROVED BY BOARD
May 11, 1990
Washington, D.C.

Roy Truby
Executive Director

May 10, 1990

The National Assessment Governing Board

Dr. Chester E. Finn, Jr., Chairman
Professor of Education and Public Policy
Vanderbilt University
Washington, D.C.

Honorable Wilhelmina Delco,
Vice-Chairman
Texas House of Representatives
Austin, Texas

Mrs. Phyllis Wil iamson Aldrich
Curriculum Coord nator
Saratoga Springs, New York

Ms. Francie Alexander
Associate Superintendent
California Department of Education
Sacramento, California

Mr. David Battini
High School History Teacher
Cairo, New York

Dr. Richard A. Boyd
Executive Director
Martha Holden Jennings Foundation
Cleveland, Ohio

Mr. Bruce E. Brombacher
Middle School Mathematics Teacher
Upper Arlington, Ohio

Honorable Michael N. Castle
Governor of Delaware
Wilmington, Delaware

Honorable Saul Cooperman
Commissioner of Education
New Jersey Department of Education
Trenton, New Jersey

Ms. Antonia Cortese
First Vice President
New York State United Teacners
Albany, New York

Mr. Christopher T. Cross, Ex Officio
Assistant Secretary
Office of Educational Research
and Improvement

Dr. Victor E. Ferry
Principal
Southwest School
Waterford, Connecticut

Mr. Michael S. Glode
Wyoming State Board of Education
Saratoga, Wyoming

Mr. Dale E. Graham
High School Principal (retired)
Columbia City, Indiana

Mr. Elton Jolly
President and Chief Executive Officer
OICs of America
Philadelphia, Pennsylvania

Dr. Carl J. Moser
Acting Director of Elementary
and Secondary Schools
Lutheran Church-Missouri Synod
Saint Louis, Missouri

Dr. Roy Truby
Executive Director, NAGB
Washington, D.C.

Mr. Mark Musick
President
Southern Regional Education Board
Atlanta, Georgia

Honorable Carolyn Pollan
Arkansas House of Representatives
Fort Smith, Arkansas

Dr. Matthew W. Prophet, Jr.
Superintendent of Schools
Portland, Oregon

Dr. Dorothy K. Rich
President
Home and School Institute
Washington D.C.

Honorable Richard W. Riley
Former Governor of South Carolina
Columbia, South Carolina

Mr. Thomas Topuses
Law Offices of Frank Rogozienski
Coronado, California

Dr. Herbert J. Walberg
Professor of Education
University of Indiana
Chicago, Illinois

TABLE OF CONTENTS

EXECUTIVE SUMMARY AND BOARD ACTION

APPROVED UNANIMOUSLY May 11, 1990
At Meeting in Washington, D.C.

Setting appropriate achievement levels on the National Assessment of Educational Progress will help define some of the important outcomes of education, stating clearly what students should know and be able to do at key grades in school. This will make the Assessment far more useful to parents and policymakers as a measure of performance in American schools and perhaps as an inducement to higher achievement. The achievement levels will be used for reporting NAEP results in a way which greatly increases their value to the American public.

The National Assessment Governing Board notes its statutory responsibility to (1) take "appropriate actions...to improve the form and use of the National Assessment" and (2) identify "appropriate achievement goals for each...grade (and) subject area to be tested under the National Assessment." To carry out these responsibilities the Board shall establish appropriate achievement levels on the National Assessment and endorses in concept the accompanying Committee paper titled, Setting Appropriate Achievement Levels for the National Assessment of Educational Progress, dated May 10, 1990. Further, the Board approves the following policy framework, definitions, and technical procedures for establishing achievement levels on the National Assessment:

1. Three achievement levels with clear distinctions between them shall be established for each grade and subject tested under NAEP. These levels shall be called:

(a) **Proficient**. This central level represents solid academic performance for each grade tested--4, 8, and 12. It will reflect a consensus that students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. At grade 12 the proficient level 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.

(b) **Advanced**. This higher level signifies superior performance beyond proficient grade-level mastery at grades 4, 8, and 12. For 12th grade the advanced level will show readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement. As data become available, it may be based in part on international comparisons of academic achievement and may also be related to Advanced Placement and other college placement exams.

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

2. It is the Board's intention to use this framework of basic, proficient, and advanced achievement levels as the primary means of reporting results for all newly-developed assessments in 1992 and thereafter. The framework shall first be applied in reporting the 1990 National Assessment of mathematics, contingent upon the successful conduct of the process to set achievement levels adopted by the Board. If the process is carried out successfully, results in terms of three achievement levels per grade shall be a prominent part of the initial release of national data from the 1990 math assessment. In the simultaneous release of data from the trial state assessment of 8th grade math, each state will have the option of having its results displayed in terms of the three achievement levels in addition to the previously-developed formats of five across-grade distributional proficiency levels, quartiles, and percent of correct answers. With the assistance of the states, the several ways of reporting results from the trial state assessment shall be evaluated.

3. The process for determining achievement levels shall be a logical continuation of the national consensus effort used in developing the content and objectives of the National Assessment.

4. To assist in defining achievement levels for the 1990 assessment of mathematics the Board shall appoint an ad hoc advisory panel, divided into separate subcommittees for grades 4, 8 and 12. The panel will be broadly representative and will consist of state and local educators, scholars, employers, civic group representatives, and other interested citizens.

5. The subcommittees will be charged with using a proven judgment procedure to recommend which test questions and/or which proportion of questions students need to answer correctly to reach various achievement levels in accordance with this framework. As part of its deliberations, the panel will be required to prepare detailed descriptions of the subject-matter knowledge and skills proposed for each achievement level. These shall be illustrated by representative sample items and scoring protocols.

6. In preparing descriptions of achievement levels and assigning test items to them the panel members shall use their best judgment and expertise and shall also take into account a wide range of background information and frames of reference. These may include relevant curriculum and testing data from state, local, national, and international levels; comments solicited from interested citizens, specialists, and education agencies; research

on the performance of different groups, such as college students and other young adults; or studies equating NAEP with other testing programs. Specifically, the panel may consider data from the 1988 International Assessment of Mathematics and Science and from Advanced Placement examinations. The panel shall refer to sources such as these in presenting the rationale for the proposed achievement levels. The panel shall ensure coherence and consistency in the recommended achievement levels over the three grades.

7. The panel shall submit proposed descriptions of mathematics achievement levels to the Board by September 20, 1990. Its report shall include sample questions, justification for the levels proposed, and a full explanation of its procedures.

8. The Board shall seek public comment on the panel's recommendations and shall hold a public forum on them during October 1990. The Board's schedule calls for it to take action on the mathematics achievement levels during its meeting of November 16 and 17.

9. It is the Board's intention that both state and national data for the 1992 assessments shall be reported initially and primarily in terms of achievement levels and that this shall be made known to the states as an element of the 1992 trial state assessment. The Board's process for establishing achievement levels will be revised as necessary on the basis of experience and practicality.

10. The Board shall ensure that all newly-developed NAEP assessments contain a broad range of content so that three achievement levels can be established for each grade in accordance with Board policy. In addition, the consensus process for developing objectives and specifications for any future assessment shall consider the three achievement levels per grade and the possibility of grade-specific scales.

11. The 1990 assessments shall continue the practice of reporting NAEP data for each subject on a common across-grade scale that spans grades, 4, 8, and 12. However, the Board is concerned that such scaling may not adequately show variations of performance within each grade. The Board intends to continue to explore the issue of grade-specific and across-grade scales. It intends to reach a decision on which scale or scales shall be used for reporting the 1992 and subsequent assessments. A timeline for making this decision shall be developed by NAGB staff, in consultation with NCES and ETS, for consideration by the Board at its August 1990 meeting.

PART 1

POLICY FRAMEWORK

BACKGROUND AND RATIONALE

Among the most significant responsibilities of the National Assessment Governing Board are (1) "taking appropriate actions... to improve the form and use of the National Assessment" and (2) setting "appropriate achievement goals" for each grade and subject tested under NAEP. The two responsibilities fit well together. By defining levels of appropriate achievement on the National Assessment the Board will increase greatly the significance and usefulness of NAEP results to educators, policymakers, and the American public.

The statute (P.L. 100-297) creating the Board assigns to it certain explicit responsibilities:

- "Taking appropriate actions needed to improve the form and use of the National Assessment;
- "Developing...standards for analysis plans and for reporting and disseminating (NAEP) results;
- "Developing standards and procedures for interstate, regional, and national comparisons;
- "Identifying appropriate achievement goals for each age and grade in each subject area to be tested under the National Assessment;
- "Developing assessment objectives (and) specifications;"
- Devising goal statements for each learning area assessment "through a national consensus approach that provides for the active participation of teachers, curriculum specialists, local school administrators, parents, and concerned members of the general public."

The National Assessment Governing Board is not authorized to establish any overarching national goals for education. It does have authority to define levels of achievement that will serve as "appropriate achievement goals" on National Assessment exams. With such achievement levels defined, NAEP results will be reported in terms that better denote the quality or value of student achievement than do the numerical scores that represent the range of student performance.

By law, the National Assessment is a survey--not a mass individual testing program--in which representative samples of students are asked questions in different academic subjects. The assessment provides information on aggregate or group performance; it is forbidden by law to report data on individuals.

Hence, the achievement levels defined by the Board will be used for reporting group data and making it more meaningful. The assessment will not become a device for certifying or classifying individual students.

In a letter to the Governing Board, Education Secretary Lauro F. Cavazos said that by "setting achievement standards for the National Assessment" the Board "would fulfill (its) statutory responsibility...(under) the Hawkins-Stafford Amendments of 1988...The result would be a clear definition of what constitutes grade level performance in each subject so that future National Assessment of Educational Progress (NAEP) reports could provide data on the proportion of students who achieve that standard and in what ways American students exceed or fall short."

The Secretary concluded that such Board action "is not only in keeping with the charge of the law, but is a constructive and complementary addition...to the work of the President and the Governors as they establish goals for performance of the Nation's education system." (Cavazos letter of Jan. 24, 1990)

THE CHANGING ENVIRONMENT

When the U.S. Office of Education was created in 1867, Congress charged it with the duty of "collecting such statistics and facts as shall show the condition and progress of education in the several states." Over the ensuing century the Office collected a great deal of information about school attendance, spending, class size, and graduates; it reported virtually nothing about what students had learned.

It was not until the mid-1960s that President Johnson and U.S. Commissioner of Education Francis Keppel sought to close this major gap by proposing a National Assessment of Educational Progress to provide data on the quality of learning in the Nation's schools. There was considerable opposition on grounds that the assessment would lead to federal control of education and a national curriculum. Similar opposition greeted the Elementary and Secondary Education Act, also proposed by Johnson and Keppel, which had as its centerpiece Title I to aid low-income students. That law passed in 1965.

The National Assessment, though, was not launched until 1969. It emerged in a form that assuaged the fears of its critics but

severely restricted its public impact and significance.

In recent years, though, the tide of opinion has turned. The U.S. Department of Education was established under President Carter in 1979. In 1983, the National Commission on Excellence in Education, appointed by Education Secretary T. H. Bell, issued its report, "A Nation at Risk." The commission somberly documented "a rising tide of mediocrity" in American schools and summoned a national movement for education reform. Bell also issued the first "wall chart" using data from Scholastic Aptitude Tests (SAT) and the American College Testing (ACT) Program to compare academic achievement in the 50 states.

Meanwhile, statewide testing programs proliferated. Almost all made public district-by-district and school-by-school comparative data. Many set standards of expected performance.

In 1988 NAEP was authorized to conduct voluntary state-by-state assessments in eighth grade math in 1990 and in fourth and eighth grade math and fourth grade reading in 1992. The same legislation created the Governing Board as an independent policy-making body for NAEP and authorized it to improve the "form and use of the assessment and to set "appropriate achievement goals."

During the past year the issue of national education goals has come to the forefront at the Charlottesville Summit of President Bush and the Nation's governors and in subsequent actions by the President and the National Governors' Association.

The need for national goals and standards was stated clearly by the Southern Regional Education Board in its 1988 report, Goals for Education:

> "If excellence means anything at all, it is a universal concept...We must be measured against the same criteria of excellence which are applied everywhere...That bold claim was controversial when made by the Southern Regional Education Board nearly three decades ago...Today, there is wide agreement that SREB states should strive for national standards. And some, particularly governors, assert that international standards are more appropriate now that the marketplace is increasingly global."

As Ernest Boyer, president of the Carnegie Foundation for the Advancement of Teaching, has declared, "The failure to establish understandable criteria and standards (for educational assessment) will lead to loss of confidence and a huge erosion of public support for the Nation's schools. We (must) give the public some evidence that our schools are working and that our $180 billion investment is paying off."

"We are now trying to...develop (national) criteria by which the performance of education can be assessed," Boyer continued, "while at the same time we retain vitality at the local level... If we could get standards straight, then we give schools some yardsticks by which they would be measured, and then we should give them a lot freedom to get there."

Setting appropriate achievement levels on the National Assessment is a step in that direction.

THE NEED FOR APPROPRIATE ACHIEVEMENT LEVELS

For the past 20 years the National Assessment of Educational Progress, like virtually all nationally standardized tests in the United States, has reported results in terms of average performance. Sometimes it has announced what proportion of students knew a certain fact or could demonstrate a certain skill. But it has shied away from saying clearly whether average performance was good enough or whether the facts and competencies it tested were ones that students really ought to know.

Of course, the NAEP assessments, like other tests, implicitly do contain judgments of significance and expected performance. Why test anything unless somebody thinks it's important? In developing NAEP, there has long been an elaborate consensus process, involving teachers, university professors, and interested groups, to determine rather precisely what body of knowledge and skills each test should measure. But again, the tests themselves and the committees creating them have only implicitly provided a basis to say how good is good enough.

As the National Academy of Science said in a report (1982), NAEP "was conceived as a white paper on the status of education in America." Its primary purpose is to report to the public on the quality of learning in the schools. But until now, the significance of its findings has often been unclear.

In an effort to improve reporting, NAEP in recent years has said what proportion of students in different grades reach different proficiency levels, but these levels--200, 250, 300, etc.--have been derived from the distribution of test results themselves, not from any prior judgment of what students ought to know. Each 50 points up or down represents one standard deviation, a measure of variation in test scores. The cluster of skills that differentiates each major level is determined by looking at the patterns of right and wrong answers after the results are in.

While helpful, such proficiency levels, are in truth simply statistical distributions. They provide limited guidance for determining whether students have mastered a challenging curriculum or have acquired the knowledge and skills needed to advance in

school or move on successfully to college and adulthood.

Defining what performance ought to be--and providing strong justification for the judgment used in making these definitions will greatly enhance NAEP's central function as a yardstick of educational achievement.

FRAMEWORK AND DEFINITIONS

The Committee recommends that the Governing Board adopt a framework for setting appropriate achievement levels that includes three levels of achievement for each grade and subject on NAEP.

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

There will be one higher level, called **Advanced**, signifying superior performance beyond proficient grade-level mastery at grades 4, 8, and 12. For 12th grade the advanced level will show readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement. As data become available, it may be based in part on international comparisons of academic achievement and may also be related to Advanced Placement and other college placement exams.

There will be one level below proficient, called **Basic**, denoting partial mastery of the knowledge and skills that are fundamental for proficient work at each grade--4, 8, and 12. For 12th grade this will be higher than minimum competency skills (which normally are taught in elementary and junior high schools) and will cover significant elements of standard high school-level work.

The Board will ensure that the content of each subject-matter assessment supports three achievement levels at each grade with clear distinctions between them. It will encourage research to permit use of international data in defining achievement levels.

This framework, applied through a broad consensus process to specific subjects in the National Assessment, will provide meaningful benchmarks of academic achievement. However, unlike any single measuring point for each grade, it will also show a wide distribution of student performance.

These benchmarks will permit states and the nation to see what proportion of students have reached very high levels of achievement on NAEP exams; strong, acceptable levels; and levels of partial mastery. Thus, it will provide a measure and incentive to improve the learning of all segments of the distribution--bottom, middle, and top.

The framework of three achievement levels at each grade is not a warrant for tracking. Indeed, the NAEP tests and the achievement levels based on them will help to ensure that all students attain competency in challenging subject matter.

The proposed achievement levels will define levels of learning tied to a common core of knowledge and skills that ought to be available to all students, regardless of family income, ethnic background, region, or type of community. The achievement goals on the National Assessment will serve to underscore the point that American schools ought not to water down what they teach the poor and beef up what they offer the more affluent.

PROCEDURES FOR ESTABLISHING SPECIFIC ACHIEVEMENT LEVELS

The process for determining achievement levels should be an outgrowth of the national consensus effort used in developing the content and objectives of National Assessment exams.

For many years NAEP has reflected a broad consensus, regularly updated by representative committees, on what is important for students to learn. In each subject area different topics at different ranges of difficulty are assessed at different grades, reflecting a consensus judgment on curricular emphases and objectives.

The proposed achievement levels will add to assessment frameworks and objectives the specific definitions of basic, proficient, and advanced achievement at each grade tested, which are based on the content of National Assessment exams. These are not broad general goals of education or curriculum, but substantive descriptions of levels of achievement tied firmly to National Assessment questions and objectives.

To assist in setting achievement levels for specific subject areas the Board will appoint ad hoc advisory panels. These will consist of state and local educators, scholars, employers, civic group representatives, and other interested citizens. The panels will be charged with using a proven judgment procedure to recommend which test questions and/or which proportion of questions students need to answer correctly to reach different achievement levels.

As part of this process, the panels will be required to prepare detailed descriptions of the subject-matter knowledge and skills proposed for each achievement level. These definitions will be based on the general descriptions adopted by the Board and will be accompanied by an explanation and rationale for the definitions proposed. It is important that there be a clear distinction between each proposed level.

The definitions of achievement levels will be similar (though presented in more detail) to the descriptions of NAEP proficiency levels prepared since 1985 by Educational Testing Service, the NAEP contractor. But, unlike the previous proficiency levels, the descriptions of achievement levels will be based on an informed, coherent judgment of what students ought to know rather than on the distribution of test results.

In preparing descriptions of achievement levels and assigning test items to them the panels should not only use their own judgment and expertise but should take into account a wide range of background information and frames of reference. These may include relevant curriculum and testing data from state, local, national, and international levels; comments solicited from interested citizens, specialists, and education agencies; research on the performance of different groups, such as literate young adults; or studies equating NAEP to Advanced Placement, Armed Forces, business, and other testing programs.

The advisory panels should refer to at least some of these sources or others in presenting and justifying their proposed definitions of achievement levels.

To illustrate the content of each proposed level, the panels --with staff assistance--will provide representative sample test items, similar to the illustrative items that have regularly been published in NAEP objectives booklets and reports. These will be accompanied by correct answers for multiple-choice items and scoring protocols for any essay or other open-ended questions.

The proposed definitions, illustrated by sample questions, will be submitted to the Board for approval. The Board will seek wide public comment before acting on the panels' recommendations.

REPORTING NAEP IN TERMS OF ACHIEVEMENT LEVELS

After appropriate achievement levels are approved by the Board and the questions and/or proportion of questions that students must answer to attain them are determined, the levels will be placed on the NAEP scoring scales. The proportion of students attaining each level will be reported.

The three achievement levels developed for each grade will be mapped onto an achievement scale. These levels will become the primary means for reporting NAEP results. However, scores at each quartile will also be reported as another means of showing the distribution of performance.

There may be advantages in using separate scales for each of the three grades in NAEP as this may be a more meaningful and educationally significant way to present assessment results. Such scales may show more clearly the variations in performance for each grade and subject in the assessment.

The scale for each grade--with basic, proficient, and advanced achievement levels clearly defined--would be distinct from any subscales for particular skills. It may be distinct from any common cross-grade scales, spanning grades 4, 8, and 12.

Under current practice, initiated six years ago, all NAEP data for each subject, such as reading or mathematics, are reported on a common scale that spans grades 4, 8, and 12. These subject-matter scales have a uniform mean score of 250, based on the performance of students in all three grades tested. Each 50 points represents one standard deviation across all students in all three grades. Because the same scale applies to grades 4, 8, and 12 the variations for each grade and subject tend to be small, especially for grades 4 and 8. 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.

Once well-developed achievement levels are established, it is the National Assessment Governing Board's intent that the stability of the achievement levels be maintained over a period of several years, perhaps a decade. Test items may be updated and the test framework may even be changed, but priority will be given to maintaining the stability of the achievement levels.

If the three-achievement level format for reporting is successfully developed, this will provide more detailed information for each grade level. Even though variations in performance within each grade will be shown more clearly, it remains to be determined whether such more detailed information will overcome the perceived shortcomings of NAEP's across-grade scale. The Board will pursue this unanswered question as it relates to the assessments of 1992 and subsequent years on a timeline to be developed by Board staff in consultation with staff of the National Center for Education Statistics and the Educational Testing Service.

WHEN SHOULD ACHIEVEMENT LEVELS BE SET?

The Committee recommends that the Board adopt the proposed framework and procedures for establishing appropriate achievement levels as policy for all future NAEP assessments. It should begin setting achievement levels with the 1990 assessment of mathematics.

The mathematics assessment is well-suited for setting appropriate achievement levels. It has been thoroughly revised through an extensive consensus process, conducted by the Council of Chief State School Officers, and incorporates many elements recommended by the National Council of Teachers of Mathematics. The assessment includes a progression of challenging topics that goes well beyond the level of basic skills where NAEP assessments have usually concentrated in the past.

The content and objectives of the math assessment have won wide endorsement from mathematics educators and state education departments. The assessment involves a field where substantial consensus already exists.

If the Board approves this proposal, it should follow the timetable adopted by NAGB on March 2, 1990. The timetable provides for the Board to appoint the panels to recommend specific mathematics achievement levels by mid-September. A public hearing or forum on these recommended levels would be held in mid-October. The Board would take final action on the mathematics achievement levels at its meeting of November 16-17, 1990.

Such a timetable would permit the achievement levels to be used in the first public reporting of nationwide data on the 1990 math assessment during the summer of 1991. State-by-state results would be reported in terms of appropriate achievement levels only at the request of individual states. The states did not know that such achievement levels would be established when they agreed to participate in the assessment. However, many states may be interested in receiving this information at the same time other state-level data are released.

This first effort at setting appropriate achievement levels should be seen as provisional and subject to further refinement and change. However, it is anticipated that the achievement levels defined will remain in place when the mathematics assessment is repeated in 1992 and for several subsequent math assessments. Soon after the math levels are set, the Board may wish to begin planning, based on that experience, to set achievement levels for the 1992 assessments of reading and writing.

NAEP AND INTERNATIONAL ACHIEVEMENT LEVELS

As the Governing Board declared in December, the National Assessment ought to become a major vehicle for comparing the achievement of American students with those of other countries. International data on student performance should be used in establishing appropriate achievement levels on NAEP exams.

The Committee proposes that the advanced level on NAEP proficiency scales become a standard of "world-class performance." As data become available, the advanced level should be based in part on high levels of performance on international assessments of student achievement.

To do this in a systematic way data would have to be obtained by having representative samples of students in other countries take NAEP assessment items, as the Board proposed in December. Alternatively, some form of equating of NAEP and other tests given internationally would be required. Some international anchoring could begin with data already available from studies conducted by the International Association for the Evaluation of Educational Achievement (IEA).

A special study was conducted in 1988 by Educational Testing Service as the first International Assessment of Mathematics and Science. In this study math and science items from the 1986 NAEP were administered to samples of 13-year-olds (mostly eighth graders) in five countries and six provincial Canadian school systems.

The proposed advisory panels to set achievement levels for math should consider these data in defining the advanced level for 8th graders on the 1990 NAEP math assessment. This might serve as an important prototype for using international data in establishing achievement levels on NAEP exams and will be helpful in determining what similar data should be obtained in the future.

REJECTED ALTERNATIVE PROPOSALS TO USE NAEP FOR SETTING ACHIEVEMENT GOALS

Two alternative suggestions have been made for setting achievement goals on the National Assessment in contrast to the appropriate achievement levels proposed in this paper. Both have serious drawbacks, as noted below. The proposals, with comment, are as follows:

1. **Use the existing NAEP proficiency levels and set targets on them for the proportion of students that should reach different levels.**

The fundamental problem with this suggestion is that the proficiency levels are not based on content but on score distributions. They are determined only after the tests are given with 250 as the mean and each 50 points representing one standard deviation. Since the scales change when NAEP tests change, previous results are sometimes recomputed, according to scales developed from the most recent testing.

In 1990 and 1992 ETS plans to give two different versions of the NAEP to two separate national samples in reading, mathematics, and writing. One version, a copy of old tests, will be used for trend data. The second version, much revised in each subject, will be used for the major cross-sectional reports and for the state-by-state assessments in math and reading. For 1994 the NAEP science test is planned to undergo a major revision through the national consensus process.

Targets might be set on the previous NAEP tests, but these would provide no data on individual states. Further, the older tests (those administered prior to 1990) have the additional drawback that much of the material on them is regarded by experts as outdated or inadequate.

Of course, goals might be set on proficiency levels that ETS establishes for the new NAEP exams. But that can't be done until the tests themselves are scored and scaled and the new levels are created. It is only at that point that anyone will know what knowledge and skills are represented by any particular level and how any level might relate to grade-level learning in school.

At that point, of course, we will know the proportion of students at each proficiency level. Any goal-setting effort would be empty unless it is for the next administration of the test, which will delay the whole process several years more.

There are three more problems with this alternative:

(a) For each subject there are only four or five defined proficiency levels, spanning all three grades tested--4, 8, and 12. This may well be too few for meaningful reporting and to show a distribution of performance at each grade. By contrast, the Committee has proposed nine levels over the same three grades.

(b) As previous data published by NAEP indicate, some of these levels have very little fit with material commonly taught at particular grade levels. Thus, they can say very little about what students have learned.

(c) Choosing what percentage of students ought to perform at a particular level is an arbitrary, poorly-defined exercise. If 5 percent of students are at a certain high level now, should 10 percent reach there in the year 2000? or 8 percent? or 12 percent? or 20 percent? Why??

We believe there is no reasonable basis for the Governing Board to set such targets. Also, there is no statutory warrant for it to try or to attempt to devise a process for doing so.

Setting targets for performance by stating what percentage of students should reach different levels is essentially a judgment that ought to be made by educational and public officials. Defining levels of performance that may serve as appropriate achievement goals on NAEP is a proper activity for NAEP's Governing Board. Others may then use the levels NAGB defines as part of their own goal-setting activities.

2. Report scores by quartiles and set targets for score increases at each quartile point.

This proposal would encounter the same problems in target-setting as the one above. There is no clear basis for setting such targets and NAGB has no warrant and no particular competence to do so. There is the further problem that no targets would be meaningful unless they were for a test that has been used in the past; both the reading and mathematics tests for the 1990 and 1992 state-by-state assessments are new, vastly different (and we think better) exams, which may not equate to previous National Assessments. The science exam may undergo major change for 1994.

Also, the point values that might be reported for each quartile have very little meaning in themselves and little significance to the public. There simply is no clear definition of the meaning of 265.8--the point value of the bottom quartile for 17-year-olds in the 1988 NAEP reading assessment. If the quartile score went up to 270, that would say virtually nothing about what additional skills or knowledge students might have. By contrast, achievement levels can be defined clearly in terms of what students know and are able to do.

Reporting by quartiles certainly is valuable for making comparisons among groups, showing the distribution of performance, and charting trends. It should continue to be part of the regular NAEP reports and should be given more prominence than it has had in NAEP reports of the past, which often have focussed on averages. However, achievement levels are a much more meaningful measure for understanding the National Assessment; these should become the principal means for reporting NAEP results.

ANOTHER SUGGESTION

It has also been suggested that NAGB not set any achievement goals or targets, but rather should devise a process that others might use to set targets for increasing the proportion of students at high levels on NAEP exams.

As discussed under alternative one above, there is no method for setting such targets which is not fundamentally an exercise in estimation and exhortation.

ENDNOTE: THE PROMISE AND SOME CAUTIONS

Setting appropriate achievement levels on the National Assessment will help define important outcomes of education, stating clearly what students should know and be able to do at key grades in school. This will make the Assessment far more useful to parents and policymakers as a measure of performance of American education and perhaps as an inducement to higher achievement.

As the National Commission on Excellence in Education noted in 1983, it is the nation that is "at risk," not just a few states. It is the whole country that is competing against the nations of Europe and Asia that today are challenging our economic position. In a Gallup poll last September over 70 percent of Americans said they favored "national achievement standards and goals."

Certainly, the Governing Board has no power of command over schools, nor does it seek such authority. NAEP hires no teachers, selects no textbooks, assigns no homework, determines no course requirements, and awards no diplomas. These are decisions made locally and by the states. The states and local governments retain full authority over what is taught in their schools. Even participation in NAEP is completely voluntary and should remain so.

However, by setting appropriate achievement levels through a broad consensus process the Governing Board has an opportunity to define a common core of learning that is important for all American children to acquire. The achievement levels will be benchmarks, points for judgment and encouragement, not edicts or commands.

If they are set well, the achievement levels will increase greatly the significance and meaning of NAEP results. Any further impact they may have will be through a process of persuasion and voluntary acceptance.

PART 2

TECHNICAL PROCEDURES

INTRODUCTION

The technology for setting achievement levels[1] has been developing over the past 35 years, and is now considered standard operating procedures for many assessment programs at the state and district level.

The technology for setting achievement levels falls into two broad categories: judgmental and empirical. Judgment methods employ appropriate groups of judges to rate the individual items in an assessment on specific criteria related to examinees' mastery or nonmastery of the content. Empirical methods use data collected from various examinee populations to make decisions about cutting scores which discriminate between two or more proficiency levels in the population. The Contrasting Groups procedure is an example of this methodology. In this approach, data from two examinee groups who clearly differ in their achievement level on the assessment are used, and the cut score is placed to maximize the discrimination between these two groups.

Judgment methods can be implemented prior to test administration, since only the items and not item data are required. However, it is highly recommended that item data, including, but not limited to, item characteristic data and distractor analysis, be made available to the panels. It is argued that allowing judges to reconsider their initial ratings and to modify those judgments generally produces more reasonable achievement levels, and reduces variability in the estimates. Item data for the 1990 mathematics assessment would be available in the late summer, and should be used by the panels in this case.

Empirical methods require that a trial assessment be administered before setting the achievement levels. It is recommended that empirical validation procedures be mounted subsequent to establishing achievement levels. Validity studies are essential in order for the achievement levels to withstand the scrutiny of the educational, business, and public sectors. It is also recommended that external validation studies be conducted where

---

[1] In this section of the staff paper the term **achievement levels** continues to be used in order to be consistent with Part 1, even though the literature has typically discussed this methodology in other terms such as standards or performance standards.

NAGB could compare the classification of groups of students according to the NAEP levels with their classification by a variety of external criteria. At the fourth and eighth grade the criteria would be school-related, whereas, at the twelfth grade criteria should include school-based and post-graduation outcome measures.

A MODIFIED ANGOFF PROCEDURE

While there are a number of competing judgment procedures that could be used for setting achievement levels, often times yielding different results, a modified Angoff procedure is recommended for a number of reasons. 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 setting achievement levels can begin immediately. However, since item data will be available, it should be used by the panels in this case. For all these reasons, and perhaps others not mentioned here, the Angoff methodology is clearly the methodology of choice.

The Angoff method will be modified to accommodate the fact that NAEP is not attempting to define the probability of a "minimally competent" student getting an item correct. As described in an earlier section of this paper, NAGB is defining achievement levels at three benchmarks on the scale, basic, proficient, and advanced.

ASSESSMENT CONTENT

A national consensus process is used to arrive at the content objectives of each subject assessed. The specific details of the process varies from subject to subject. However, the overall concept involves various publics in advising the Board on the current theoretical, curricula, and instructional status of any given content area. The process includes numerous iterations filtering each perspective through that of competing ones, until a final product is derived which represents the best thinking in the field and for which there is general agreement.

In the basic areas, such as reading and mathematics, and, indeed, in all the NAEP core areas, there is an underlying assumption of a developmental curriculum. That is, specific objectives span several years as the students' capacities develop from the lower levels of the content taxonomy in the elementary grades to the highest levels at the upper grades. This approach ultimately forms the conceptual basis of the NAEP scales which

currently cut across grade levels and are behaviorally anchored to real tasks and accomplishments at specific intervals on the scale.

The content objectives are then defined in measurable terms as the consensus process continues to spell out the test and item specifications. In other words, the consensus process moves toward articulating not only content expectations at each grade level, but the parameters within which those objectives will be assessed. Typically, the field testing of an item pool follows and the final selection of appropriate assessment items is made by the Board.

ACHIEVEMENT LEVELS

In identifying the content specifications for each subject area assessed, there is an underlying assumption that all students in grade 4, for example, should be able to respond to questions about the "volume of rectangular solids." In other words, this objective would not have been assigned to grade 4 if the framework had not placed it there. This is a reflection of the criterion-referenced nature of NAEP. However, due to measurement error in the assessment, and due to the less-than-perfect performance of students on the assessment, in any given grade level there will be a distribution of performance. So, even though the "ideal" expectation for grade 4 as described by the test objectives might include knowledge of the "volume of rectangular solids," a more accurate expectation for grade 4 can be derived by the careful examination of the items designed to measure the grade 4 assessment objectives.

Achieving consensus on the real expectation for students is the process of setting achievement levels, the yardstick by which the degree of success on the subject matter content for each grade will be assessed.

Setting definitive achievement levels for each grade and in each subject area assessed allows users of NAEP to make informed judgments about the quality of the results, and seeks to provide answers to the following questions: How good is good enough? Do we have substantially different expectations for different content areas? Are there levels of achievement within each content area that distinguish those who are truly proficient in the content from those who are only modestly proficient? Setting achievement levels for NAEP will assist us in answering those questions, and in interpreting the data better.

NUMBER OF LEVELS AND SCALES FOR EACH GRADE

Earlier it was mentioned that three achievement levels would

be established for each grade level. We must caution, that in order to accomplish three levels at each grade level, the distribution of item difficulty and content must be adequate (1) to support the accurate and precise description of collective examinee performance in the four achievement regions defined by the achievement levels, and (2) to describe examinees' collective abilities to perform tasks that are deemed to be clear and interpretable by educators and the public.

At the present time, with a single cross-age/grade scale, there are five benchmarks. If three unique grade scales are established, with three benchmarks each, this results in nine achievement levels, four more than NAEP now has. It is not clear at this point whether or not the data will support this increase. However, preliminary judgments seem to indicate that it should. This issue certainly will need to be reexamined for each subject area, particularly as the one hour response time for examinees is used to provide more extended responses on fewer numbers of items.

On how many scales or subscales should achievement levels be set? A sufficient number of scales should be created to represent accurately achievement on all or nearly all of the exercises in the pool at a given grade level. As many exercises as possible should be incorporated into the IRT scales. This may entail some revision of initial plans for scaling. It must be recognized, however, that small, important groups of exercises may remain, which are insufficient to support separate IRT scales but sufficiently important and substantive enough to warrant not setting aside. In such cases, item clusters may be scaled using alternate techniques. Scale scores developed by alternate methods should be expressed in metrics comparable to those used for IRT-based scales.

When more than one scale is required to represent accurately achievement on all or nearly all of the exercises, an index should be created by taking a weighted composite of scales, the weights to be determined by a rational, deliberative procedure. Whenever possible, achievement levels should be established and reported for all scales as well as the composite indices.

PROCEDURES FOR SETTING ACHIEVEMENT LEVELS

There are probably hundreds of variations on what has become known as the "Angoff Method." This is because a method for setting achievement levels includes much more than simply the nature of the judges' rating task. In developing the method to be implemented, reference and consideration must be given to the following features of the process discussed here.

Composition of the Panels

The groups to be represented on the panels must be identified,

and procedures for selecting representatives must be determined. It is recommended that the panels be composed of individuals with expertise in the education of students of the ages and grades under consideration, in the subject areas under consideration, with experience in the assessment of students' achievement in the subject areas under consideration, with knowledge of the typical subject area achievement of students of the ages and grades under consideration, and, in the case of twelfth grade assessments, with knowledge of the subject area achievement requirements of high school graduates who aspire to post-high school experiences in the work force, the military, or post-secondary education programs.

Major national organizations will be contacted to recommend from among their members individuals who might serve on the panels as well as alternates. In selecting members for the panels great care will be exercised in making certain that the required and desired demographic and technical characteristics are represented on the panels.

There are two additional criteria which must be applied when designing the composition of the panels. First, there should be some continuity with the mathematics consensus panels convened in 1988 to recommend the content and objectives of the 1990 assessment. Therefore, some members of the previous panels should be requested to serve on the panels. The second criteria must ensure that states participating in the 1990 state-by-state trial assessment be represented on the panels as well.
This is particularly important at the eighth grade level.

Size of the Panels

How many judges should there be? This is a technical issue which is not easy to answer. Generally speaking, the larger the sample of judges on the panels the less error of estimation there will be. However, every estimation procedure which employs a sample to estimate a population parameter will have some amount of error associated with it. In addition, every instrument has a margin of error associated with it called the standard error of measurement. Setting standards, therefore, does add a second source of error. It is desirable to keep this additional source of error at a minimum, so that the overall standard error is not excessively large.

It is recommended that a sufficient number of judges be on the grade level panels such that the overall standard error is increased by no more than 12%. This can be achieved by ensuring that the standard error of the mean recommended grade level achievement levels is no more than 0.5 of the standard error of measurement of the assessment. The research has suggested that this criterion will probably necessitate having between 16 and 20 judges on each grade level panel, that can be divided into four groups of 4 or 5 judges each. Each group will be chosen, if

possible, to be representative of the entire group. In that way, independent replications of setting the achievement levels process can be conducted and the resulting achievement levels compared.

Training of the Judges

It is recommended that training for the panels include training both to the task and the process. This training would include, but not be limited to, definitions of the three achievement levels, the rating method to be used, and the adjudication of extreme ratings through panel iterations. It is critical that the training include practice exercises with feedback, and several simulations to ensure full comprehension of the task, and full understanding of the definitions of the benchmarks. Of special interest will be training judges to provide multiple ratings for each item corresponding to the benchmark points of interest.

Resources Available to Judges

As discussed earlier it is highly desirable to have item characteristic data available to the judges after they have made their initial ratings of items. Allowing the panels to have the data to condition their final judgments usually leads to more reasonable and converging achievement levels. An informed panel is more apt to make sound judgments than an uniformed panel. Since in math the 1990 data will be available at or around the time the panels meet, it is in the best interest of defensible achievement levels that the panels be given such data.

In addition, judges will have the test and item specifications available, the content area framework, and all the items coded by grade and objective, and an answer key.

Briefing materials will also be prepared for the judges that will assist the panels in making a more informed judgment about the objectives and exercises in the assessment. These materials might include, but would not be limited to, a variety of supplementary documents and external criteria that could assist the judges in evaluating their individual estimates of achievement levels in each assessment.

General Meeting Strategies

Each panel member will review the framework of the assessment as well as the test and item specifications. Each judge will then be instructed in how to use the Task Review Form (or a form similar to the one shown in Appendix A). Each judge will complete the Task Review Form, and then, as a group, they will determine a consensus average percent for each objective. In reaching a consensus, the discussion will focus on outlier ratings, and each judge will have the opportunity to reconsider h/er own ratings. This procedure

will be completed three times, once for each of the three benchmarks. A final listing of ratings for each objective will be compiled, each representing a profile of the content that a group of students who meet the benchmark criteria should have mastered. These consensus ratings will be added to the Item Review Forms (or a form similar to the one shown in Appendix B).

Once the panels have had the opportunity to work with several practices exercises (items), the judges will complete the item reviews individually. Within the smaller groups of 4-5, judges will discuss their individual ratings to reach consensus. Individual judges will aggregate their own ratings to produce an individual achievement levels, and finally aggregate them to produce group achievement levels. This will be completed three times, once for each benchmark.

The smaller groups of judges will then come together to compare their group achievement levels, and to reach consensus as a panel on a single achievement level, one for each benchmark. It is at this point that empirical data from the assessment will be made available to the panels for their consideration. Should judges wish to modify their ratings before reaching a final judgment they can do so at this time.

Describing the Anchor Points

Once the panels have completed their work, the final ratings of the judges will be aligned with the items on the assessment placed in order of their scale values. This graphic representation[2] will display the location of the items on the IRT scale (if available), the degree of agreement among the panel members, and will be used by the panels to generate the content descriptions of the anchor points. Such descriptions will be accompanied by representative items for each point either from the released item pool or other items written specifically to demonstrate the content.

Documenting and Evaluating the Process

A complete record of the meetings and the process used by the panels will be made, so that problems, inconsistencies, or other

---

[2] The suggestion for a graphic display was made by Edward Haertel, Stanford University, at a meeting held in Chicago on February 24, 1990, with NAGB and ETS staff.

issues can be addressed in subsequent achievement level activities.

The Board will conduct a formal evaluation of the process. The evaluation will cover all aspects of the process, from both a technical and policy perspective, and will make recommendations for improving future activities in this area.

Appendix A

Task Review Form

Task Review Form

Strategy: This form should be used with the group of judges to help the group reach a joint understanding of what minimum competency is for each task or objective. (In the form, the word "Task" is substituted for "Sub-Responsibility" for convenience.)

Each judge should determine the percent of times that a task or objective is to be accomplished with no or only a few minor errors. As a group, the judges should reach a compromise rating among their collective ratings.

Form:

Directions: Read each task in the role delineation statement (domain specification or objective) and determine the percent of times each task (objective) must be accomplished with no or only a few minor errors. For example, consider the following task:

Complete a standard order form for ordering office supplies

For this example, what percent of items that an order form is to be completed must the form be completed with no or only a few minor errors?

Task X. _____ %

The response is _____ % of the times the order form must be completed with no or only a few minor errors.

Now, ask judges to look at the tasks in the role delineation profile.

What percent of times should each task be performed with no or only a few minor errors?

Write a percent in the space provided.

1. _____%	11. _____%	21. _____%	31. _____%
2. _____%	12. _____%	22. _____%	32. _____%
3. _____%	13. _____%	23. _____%	33. _____%
4. _____%	14. _____%	24. _____%	34. _____%
5. _____%	15. _____%	25. _____%	35. _____%
6. _____%	16. _____%	26. _____%	36. _____%
7. _____%	17. _____%	27. _____%	37. _____%
8. _____%	18. _____%	28. _____%	38. _____%
9. _____%	19. _____%	29. _____%	39. _____%
10._____%	20. _____%	30. _____%	40. _____%

Appendix B

Angoff Item Review Form
(Method A)

Angoff Item Review Form

Reviewer's Name: _____ Date: _____

Task (Objective) Statement: (insert the task objective number here)

This task objective must be performed _____% of the time with no or only a few errors.

I. Ask judges to think of a group of persons who are just able to meet this required level of performance for this task (objective). The exam items below were prepared to measure this task (objective). What percent of the group of people that you are thinking about will be able to answer each exam item correctly? Write the percent (between 0 and 100) for each exam item in the column labelled "Initial Percent."

Test Item	Initial Percent	Revised Percent
_____	_____%	_____%
_____	_____%	_____%
_____	_____%	_____%
_____	_____%	_____%
_____	_____%	_____%
_____	_____%	_____%
_____	_____%	_____%
_____	_____%	_____%
_____	_____%	_____%
_____	_____%	_____%

II. When the judges in the work group have provided their initial ratings, ask them to compare their percents on an item-by-item basis. Also, review the scoring key. Identify the judges who have the highest and lowest percent for each exam item. If they are greatly different (about 20% points difference) then they should discuss why the percents were chosen. They do not have to reach a compromise. Only reconsider their own ratings when there are large differences. If they want to change their percents for any exam item, they should write a new percent in the Revised Percent column.

PART 3

DISPLAYING NAEP RESULTS IN TERMS OF ACHIEVEMENT LEVELS

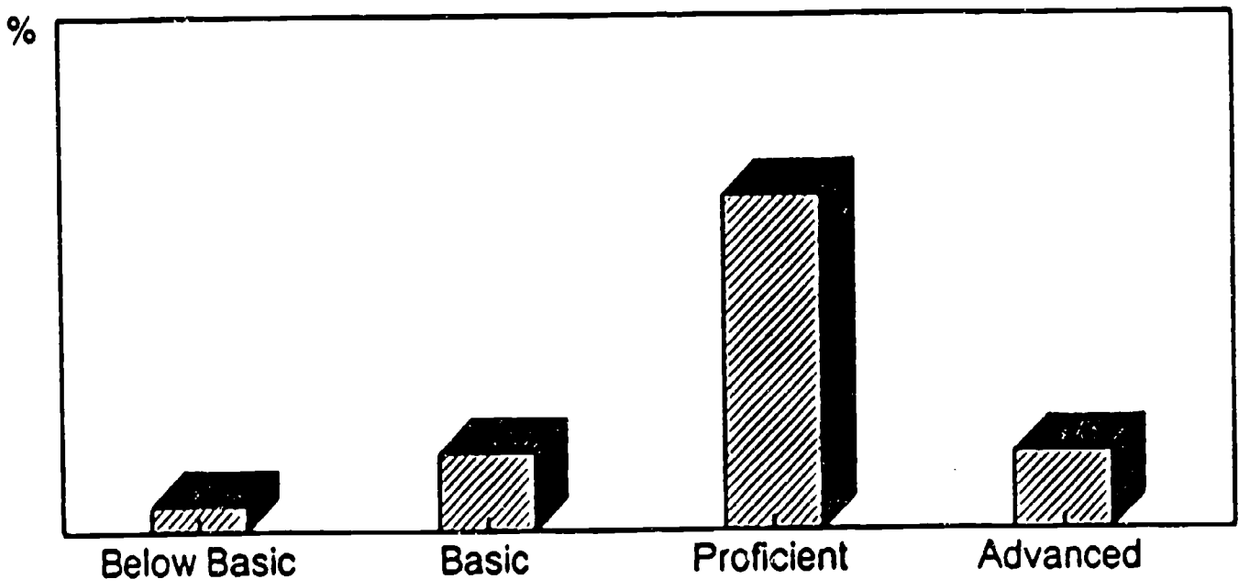
Once achievement levels have been established for a given subject area assessment, the results can be reported in terms of these levels in a variety of ways. Reports of NAEP results can be tailored to specific audiences, thereby increasing the significance and usefulness of NAEP data to educators, policymakers, and the general public.

The graphics on the following pages depict some of the many forms and formats for reporting NAEP results based on the achievement levels. The figures in Sample 1 illustrate two ways to look at performance for the distribution. For a single year, the percentage at each achievement level could be graphed as shown in the first chart. Similarly, the second chart shows changes in the percentage of students at each level over time on successive administrations of a subject area assessment.

Individual states may wish to set targets by establishing, for example, the percentage of students expected to reach each achievement level. Progress toward these targets could then be displayed, as shown in Sample 2. A value-added approach, as depicted in Sample 3, could present the progress toward a state-defined goal over time. Finally, Sample 4 illustrates the use of achievement levels to show gaps between various subgroups on the NAEP scale.

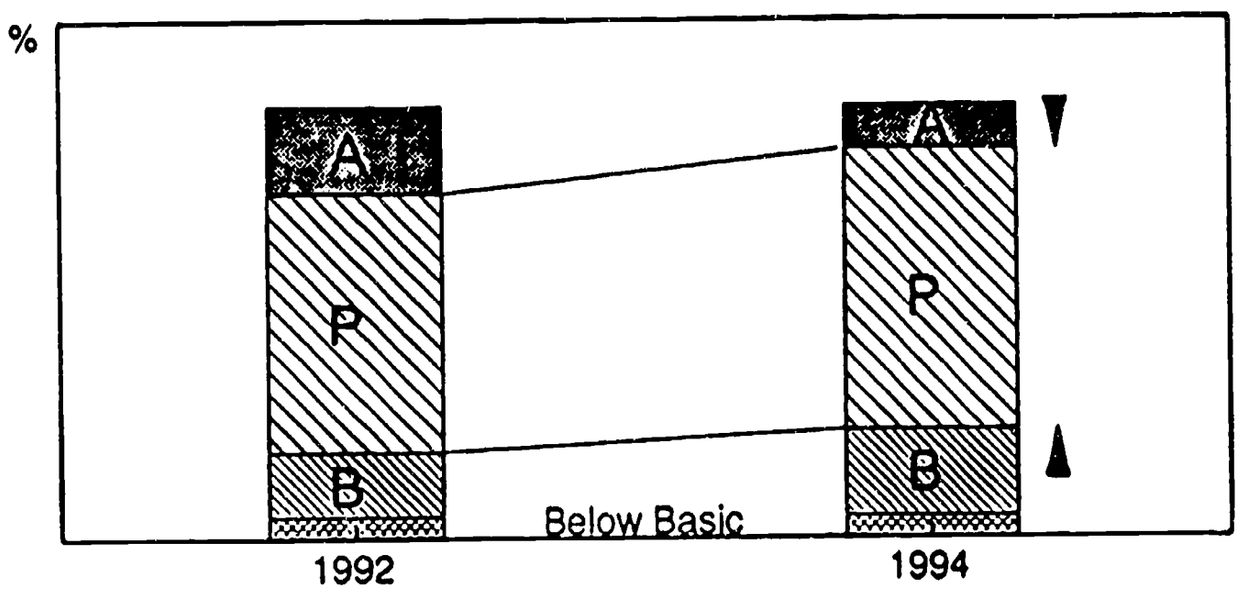
These charts, though general in nature, do serve to illustrate some of the many ways in which the NAEP achievement levels can enhance the interpretability and usefulness of the National Assessment results for diverse audiences.

PERFORMANCE FOR THE DISTRIBUTION



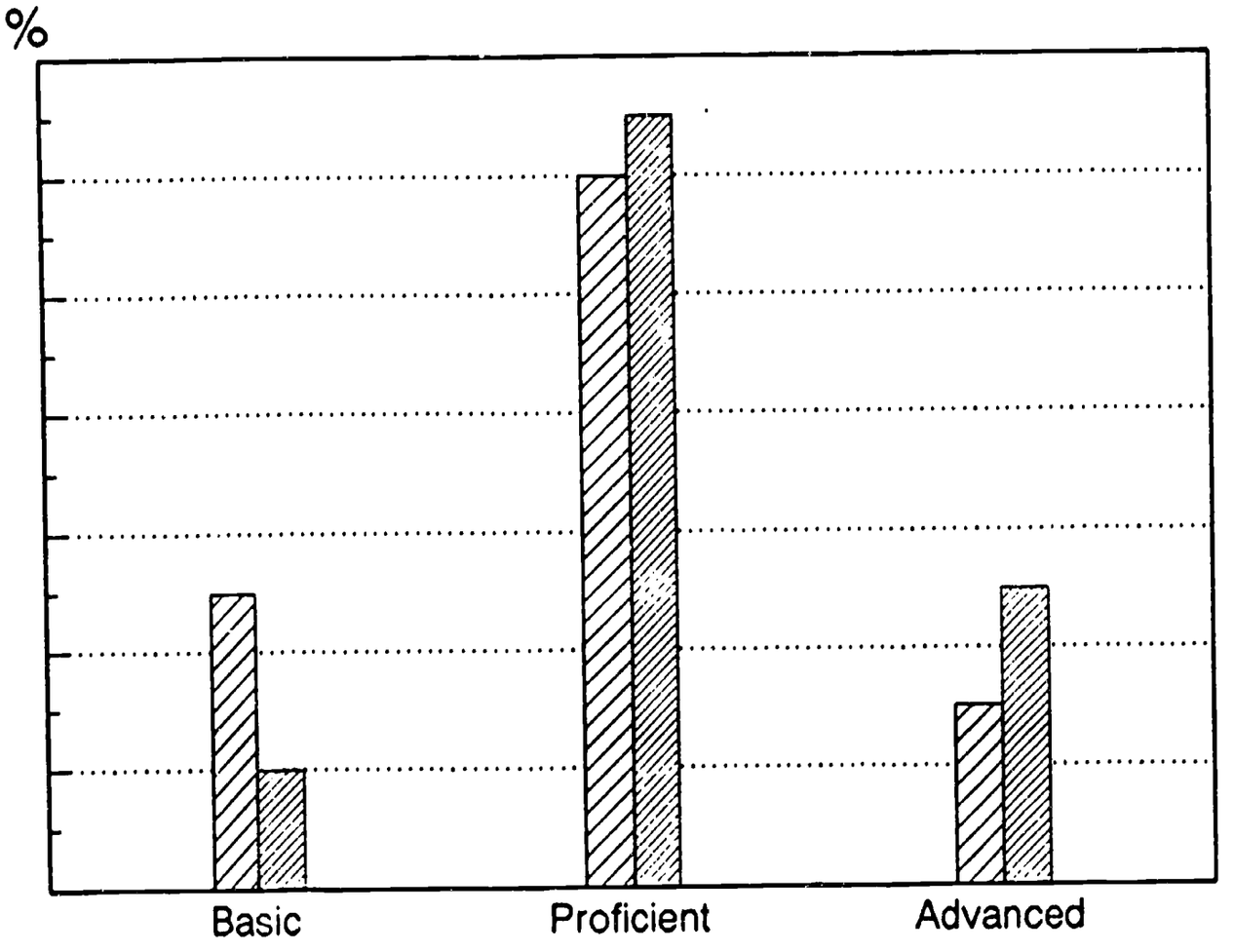
Percentage at Each Level

1992



Percentage Change at Each Level Over Time

Progress Toward Targets

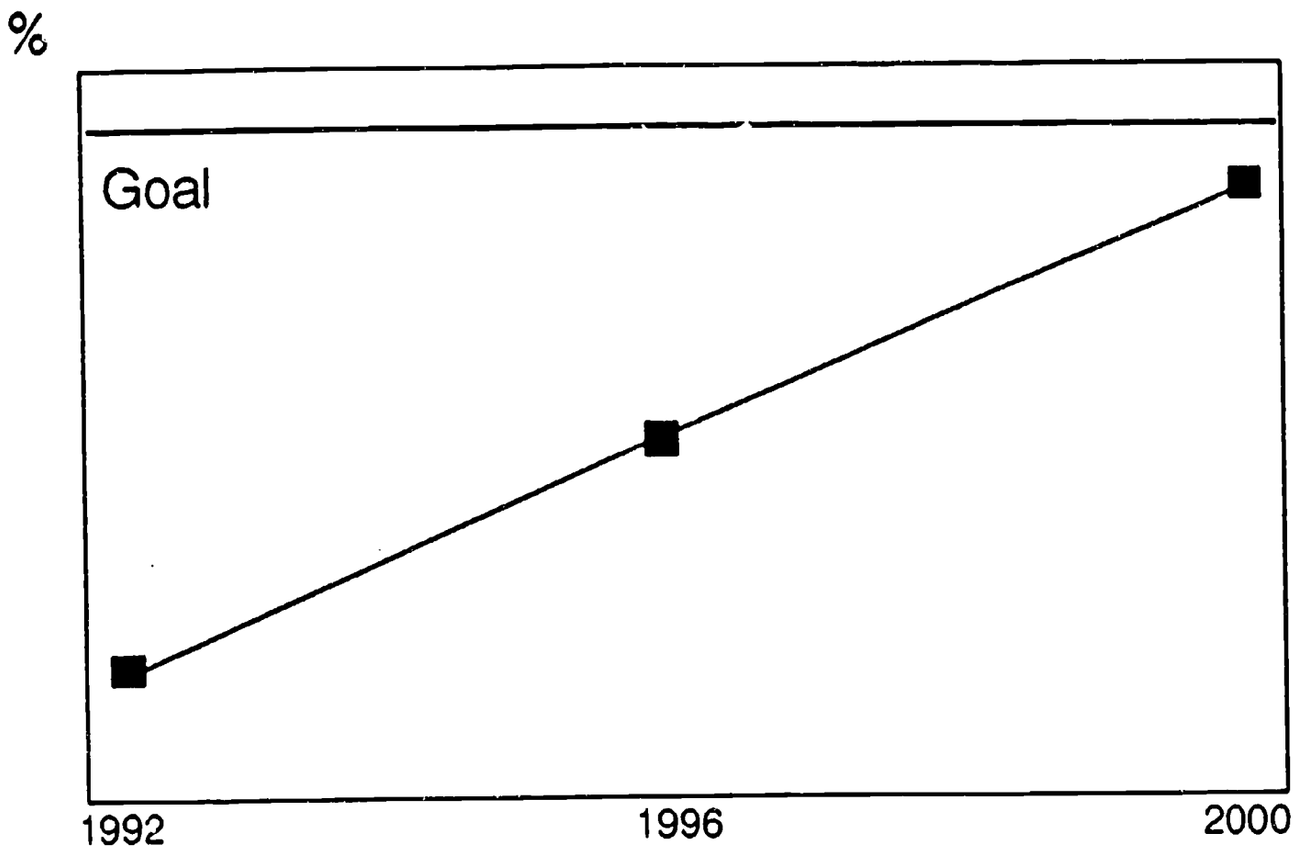


1992

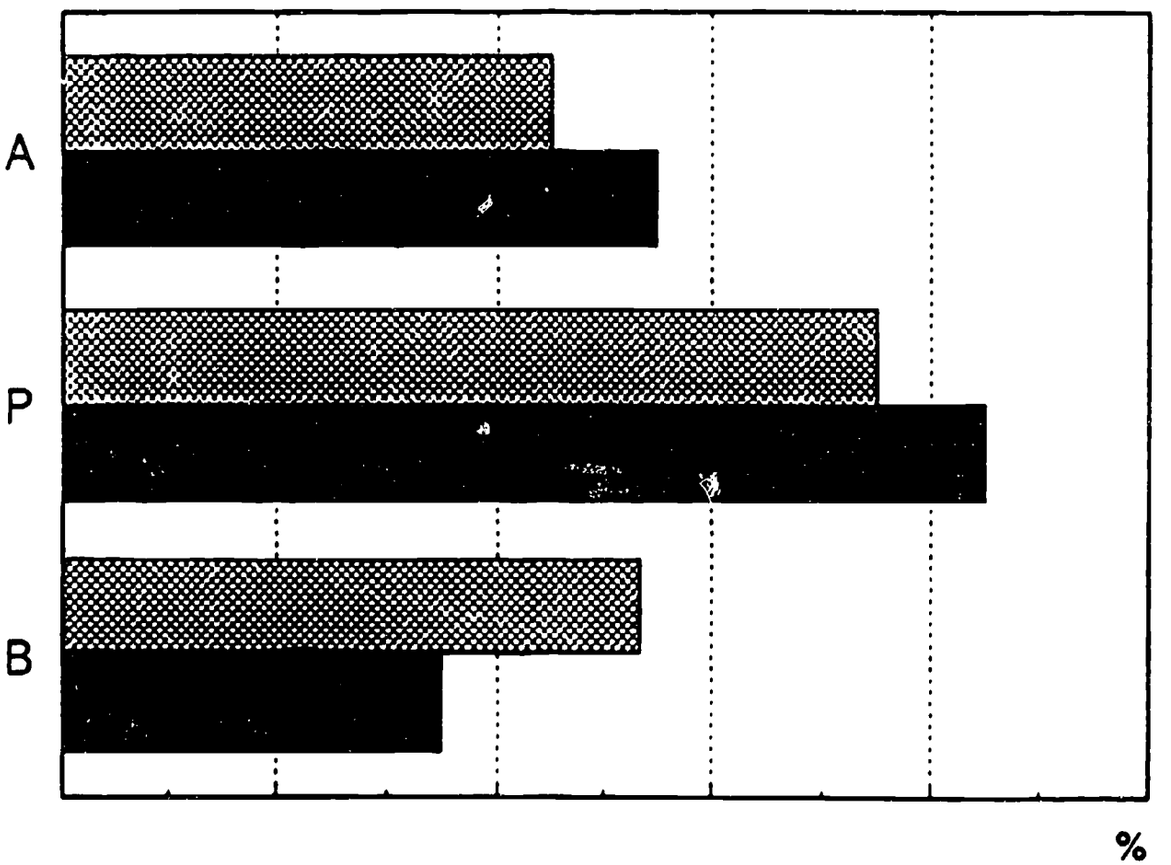
Achieved Levels

Targeted Levels

Growth Over Time - Value Added Approach



Gaps Between Subgroups



PERCENTAGE AT EACH LEVEL