

Adopted: March 4, 1995



## National Assessment Governing Board

### Developing Student Performance Levels for the National Assessment of Educational Progress

#### Policy Statement

##### *Foreword*

*A policy on setting achievement levels on the National Assessment of Educational Progress (NAEP) was first adopted in 1990 and amended several times thereafter. The present policy, adopted in 1995, contained introductory and explanatory text, principles, and guidelines. Since 1995, there have been several changes to the NAEP authorizing legislation (currently, the NAEP Authorization Act: P.L. 110-279). In addition, related legislation has been enacted, including the No Child Left Act of 2001. Consequently, introductory and other explanatory text in the original version of this policy, no longer germane, has been deleted or revised to conform to current legislation. The Principles and Guidelines remain in their original form except for Principle 4, from which the reference to the now decommissioned Advisory Council on Education Statistics has been deleted. (Foreword added August 2007.)*

#### Principles for Setting Achievement Levels

##### **Principle 1**

The level setting process shall produce for each content area, three threshold points at each grade level assessed, demarcating entry into three categories: *Basic*, *Proficient*, and *Advanced*.

*Proficient.*      *This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.*

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

## **Principle 2**

Developing achievement levels shall be a widely inclusive activity of the Board, utilizing a national consensus approach, and providing for the active participation of teachers, other educators (including curriculum specialists and school administrators at the local and state levels), and non-educators including parents, members of the general public, and specialists in the particular content area.

The development of achievement levels shall be conducted in two phases. In phase 1, the assessment framework development process shall yield preliminary descriptions of the achievement levels (*Basic, Proficient, and Advanced*), which shall subsequently be used in phase 2 to develop the numerical standards (cut scores) and to identify appropriate examples of assessment exercises that typify performance at each level. The levels will be updated as appropriate, typically when the assessment frameworks are updated.

## **Principle 3**

The Governing Board shall incorporate the student performance levels into all significant elements of NAEP, including the subject area framework development process, exercise development and selection, and the methodology of the assessment. The achievement levels shall be used to report the results of the NAEP assessments so long as such levels are reasonable, valid and informative to the public.

## **Principle 4**

In carrying out its statutory mandate, the Governing Board will *exercise its policy judgment in setting the levels*. The Board shall continually seek better means of setting achievement levels. In so doing, the Board may seek technical advice as appropriate from a variety of sources, including external evaluations provided by the Secretary, the Commissioner, and other experts. Proposed achievement levels shall be reviewed by a broad constituency, including consumers of NAEP data, such as policymakers, professional groups, the states and territories. In carrying out its responsibilities, the Board will ordinarily engage the services of a contractor who will prepare recommendations for the Board's consideration on the levels, the descriptions, and the exemplar exercises.

## **Guidelines for Setting Achievement Levels**

Each guideline presented below is accompanied by a rationale and a summary of the implementation practices and procedures to be followed in carrying out the principle. It should be understood that the full implementation of this policy will require the

contractor, through Governing Board staff, to provide assurances to the Board that all aspects of the practices and procedures for which they are responsible have been completed successfully. These assurances will be in writing, and may require supporting documentation prepared by the contractor and/or Governing Board staff.

## **Summary of Guidelines**

### **Guideline 1**

The level setting process shall produce for each content area, three threshold points at each grade level assessed, demarcating entry into three categories: *Basic, Proficient, and Advanced*.

### **Guideline 2**

The level setting process shall be a widely inclusive activity of the Board, carried out by a broadly representative body of teachers, other educators (including curriculum specialists and local and state administrators), and non-educators including parents, concerned members of the general public, and specialists in the particular content area; this process and resulting products shall be reviewed by a broad constituency.

### **Guideline 3**

The level-setting process shall result in achievement level cut scores for each grade and level, expanded descriptions of the content expected at each level based on the preliminary descriptions provided through the national consensus process, and exemplar exercises that are representative of the performance of examinees at each of the levels and of the cognitive expectations for each level described.

### **Guideline 4**

In carrying out its statutory mandate, the Board will *exercise its policy judgment in setting the levels*. However, in so doing, they will seek technical advice from a variety of sources, but especially from the contractor who will prepare the recommendations on the levels, the descriptions, and the exemplar exercises, as well as from consumers of NAEP data, including policymakers, professional groups, the states, and territories.

### **Guideline 5**

The achievement levels shall be the initial and primary means of reporting the results of the National Assessment of Educational Progress at both the national and state levels.

### **Guideline 6**

The level-setting process shall be managed in a technically sound, efficient, cost-effective manner, and shall be completed in a timely fashion.

## **Guideline 1**

The level setting process shall produce for each content area, three threshold points at each grade level assessed, demarcating entry into three categories: *Basic, Proficient, and Advanced.*

## **Rationale**

The Board is committed to describing the full range of performance on the NAEP scale, for students whose performance is in the mid-range, as well as for those whose performance is below and above the middle. It is highly desirable to endorse realistic expectations for all students to achieve no matter what their present performance might be. Three benchmarks on the NAEP scale suggest realistic expectations for students in all regions of the performance distribution. Likewise, the Board is committed to preserving trend results in NAEP. Three achievement levels accommodate growth (and possible declines) in all ranges of the performance distribution.

## **Practices and Procedures**

### **Policy Definitions**

The following policy definitions will be applied to all grades, 4, 8, and 12, and all content areas in which the levels are set. It is the Board's view that the level of performance referred to in the policy definitions is what students *should be able to know and do*, and not simply the current academic achievement of students or that which today's U.S. schools expect.

- Proficient.* This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.
- Basic.* This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.
- Advanced.* This level signifies superior performance beyond proficient.

### **From Policy Definitions to Content Descriptions**

In the course of applying the policy definitions to the level-setting process, it will be necessary to articulate them in terms of the specific content and sequence (now called descriptions) appropriate for the grades in which the levels are being set. This will be completed on a preliminary basis through the process which develops the assessment

frameworks. These preliminary descriptions will be used to initially guide the work of deriving the advice that will assist the Board in setting the levels. Throughout the process of obtaining such advice, however, these descriptions may be refined, expanded, and edited to more clearly reflect the specific advice on the levels.

### **Training of Judges**

In training the judges for the level-setting activity, it is necessary that all arrive at a common conceptualization of *Basic, Proficient, and Advanced* based on the policy definitions of the Board. Such conceptualizations must be within the scope of the assessment framework under consideration and capable of being applied at the individual item level (Reid, 1991.)

Judges must also be trained in the specific model that will be used to generate the rating data. At the very least, they need to understand the purposes for setting the levels, the significance of such an activity, the NAEP assessment framework for the subject area under discussion, elements that make particular exercises more or less difficult, and the rating task itself.

Judges shall be trained by individuals who are both knowledgeable in the subject matter area and are experienced, capable trainers in a large-group setting. Presentations shall be prepared, rehearsed, and piloted before implementation.

Judges shall be provided comprehensive, user-friendly training materials, adequate time to complete the task, and the appropriate atmosphere in which to work, one that is quiet, pleasant, and conducive to reaching the goals of the level-setting activity. It is also required that judges take the assessment under the same NAEP-like conditions as students, that is, using the NAEP student booklets, having all manipulatives and ancillary materials, and timed.

### **Guideline 2**

The level setting process shall be a widely *inclusive* activity of the Board, carried out by a broadly representative body of teachers, other educators (including curriculum specialists and local and state administrators), and non-educators including parents, concerned members of the general public, employers, scholars, and specialists in the particular content area. This process and resulting products shall be reviewed by a broad constituency.

### **Rationale**

The spirit of the legislative mandate of the Board is one of moving toward a national consensus on policy issues affecting NAEP. The Board has historically involved broad audiences in its deliberations. The achievement levels are no different. Further, the Board views the level-setting activity as an extension of the widely inclusive effort to derive the assessment frameworks and scope and sequence of each assessment. Finally, the magnitude of the decisions regarding *what students should know and be able to do is*

simply too important a decision to seek involvement from professionals alone; it must have the benefit of the collective wisdom of a broadly representative body, educators and non-educators alike.

## **Practices and Procedures**

### **Sample of Judges**

The panel of judges will be composed of both educators and non-educators. About two-thirds of the panel will represent teachers and other educators; one-third will represent the public, non-educator sector, for example, scholars, employers, parents, and professionals in occupations related to the content area. They will be drawn from a national sampling frame and will be broadly representative of various geographic regions (Northeast, Southeast, Central, West, and the territories) types of communities (urban, suburban, rural), ethnicities, and genders.

Individual panel members shall have expertise in the specific content area in which the levels are being developed, expertise in the education of students at the grades under consideration, and a general knowledge of assessment, curriculum, and student performance. The composition of the panels should be such that they meet the requirements of the *Standards (1985)*.

The size of the panels should be responsive to what the research demonstrates regarding numbers of judges involved (see Jaeger, 1991). While it may not be practical or beyond the resources available, every effort should be made to empanel a sufficient number of judges to reduce the standard error of the cut score. While there is no absolute criterion on the magnitude of the standard error of the cut score, a useful rule of thumb is that it should not exceed the *combined* error associated with the standard error of measurement on the assessment and the error due to sampling from the population of examinees.

### **Review Procedures**

Throughout the process and particularly at critical junctures, groups that have a legitimate interest in the process will be involved. During the planning process interested groups and individuals will be encouraged to participate and share their experiences in the area of setting standards. These groups might include professional societies, *ad hoc* advisory groups, standing advisory committees to the Governing Board or its contractor(s) and NCES and its contractor(s) and grantees. Documents (such as the Design Document and Interim Reports) will be disseminated in sufficient time to allow for a thoughtful response from those who wish to provide one.

Proposed levels will be widely distributed to major professional organizations, state and local assessment and curriculum personnel, business leaders, government officials, the Planning and Steering Committees of the framework development process, the Exercise Development panels, and other groups who may request them.

When it is deemed useful by the Board, public hearings and forums will be conducted in Washington, D.C. and other parts of the country to encourage review and input on a broad regional and geographic basis.

### **Guideline 3**

The resulting products of the level-setting process shall be (1) achievement level scores marking the threshold score for each grade and level, (2) expanded descriptions of the content expected at each level based on the preliminary descriptions provided through the national consensus process, and (3) exemplar exercises that are representative of the performance of examinees at each of the levels and of the cognitive expectations for each level described. These three products form the basis for reporting the results of all future NAEP assessments.

### **Rationale**

The NAEP scale, while useful for aggregating large amounts of information about student performance in a single number, requires contextual information about the specific content and the sequencing of that content across particular grades, in order to be truly beneficial to users of NAEP data. In order to make the NAEP data more useful, descriptions of each level which articulate content expectations and exemplar exercises taken from the public release pool of the most current NAEP assessment must accompany the benchmarks or cut scores for each level. The descriptions and exemplars are intended to be illustrative of the kind of content that is represented in the levels, as well as an aid in the interpretation of the NAEP data.

## **Practices and Procedures**

### **Methodology**

The methodology to be used in generating the levels will depend upon the specific assessment formats for the content area in which the levels are being set. Historically, in the case of multiple choice exercises and short constructed response formats, a modified Angoff (1971) procedure has been employed. In the case of extended constructed response formats, a paper-selection procedure has been employed. Neither of these is without its disadvantages. As the assessment formats of future assessments become more complex and employ more performance-type exercises, it is quite likely that alternate procedures will be needed. The Board will decide these on a case-by-case basis, looking for advice from those who have had experience in dealing with these alternative assessment formats. In any case, the design for carrying out the process must be carefully crafted, must be appropriate to the content area and philosophy of the assessment framework, and must have a solid research base.

The procedures will generally be piloted prior to full implementation. The purpose of the pilot would be to test out the materials used with the judges, the training procedures, the feedback information given to the judges during the process, and the

software used to complete the initial analyses. Procedures would be revised based on the pilot experience and evaluation evidence.

Whatever methodology is used, all aspects of the procedures will be documented for the purposes of providing evidence of procedural validity for the levels being recommended. This evidence will be made available to the Board at the time of deliberations about the levels being set.

### **Quality Control Procedures**

While there are numerous points in a complex process for mistakes to occur, there are at least three important junctures where quality control measures need to be in place. First, is the point of data entry. Ideally, judges' ratings should be scanned to reduce manual errors of entry. However, if the ratings are entered manually, then they shall be entered and 100% verified using a double-entry, cross-checking procedure. Second, software programs designed to complete initial analyses on the rating data must be run with simulated data to de-bug, and provide assurances of quality control. The programs should detect logical errors and other kinds of problems that could result in incorrect results being generated. Finally, the production of cut scores on the NAEP scale is the final responsibility of the NAEP operations contractor. Only final cut scores, mapped onto the properly weighted and equated scale, received in writing from the operations contractor, will be officially communicated to the Board, or others who have a legitimate need to know. *Once the accuracy of the data has been ensured by the level-setting and operations contractors, the Board shall make a policy determination and set the final achievement levels, informed by the technical process of the level-setting activity.*

### **Descriptions of the Levels**

The preliminary descriptions developed through the framework development process will be the starting point for developing recommendations for the levels under consideration. The preliminary descriptions are *working descriptions* for the panels while doing the ratings. These may be expanded and revised accordingly as these panels conduct the ratings, examine empirical performance data, and work to develop their final recommendations on the levels. The recommended descriptions will be articulated in terms of what students *should know and should be able to do*. They shall be coherent within grade, and consistent across grades, and will reference performance within the three regions created by the cut scores. No descriptions will be done for content below the *Basic* level.

### **Exemplar Exercises**

The exemplars chosen from the released pool of exercises for the current NAEP assessment will reflect as much as possible performance both in the *Basic, Proficient, and Advanced* regions of the scale, as well as at the threshold scores. Exemplars will be selected to meet the  $rp = .50$  criterion, and will demonstrate the range of performance possible within the regions. They will likewise reflect the content found in the final descriptions and the range of item formats on the assessment. Evidence will be provided for the degree of congruence between the content of the exemplars and that of the descriptions. There will be at least three exemplars per level per grade identified.

## **Guideline 4**

In carrying out its statutory mandate, the Board will *exercise its policy judgment in setting the levels*. However, in so doing, they will seek technical advice from a variety of sources, but especially from the contractor, who will prepare the recommendations on the levels, the descriptions, and the exemplar exercises, as well as from consumers of NAEP data, including policymakers, professional groups, the states and territories.

## **Rationale**

Setting achievement levels is both an *art* and a *science*. As an *art*, it requires judgment. It is the Board's best policy judgment what the levels should be. However, as a *science*, it requires solid technical advice based on a sound technical process. The Board is committed to seeking such technical advice from a variety of sources.

## **Practices and Procedures**

### **Technical Advice throughout the Process**

The Board seeks to involve persons who have had experience in standard-setting at the state level, and from those who are users of the NAEP results. Regular presentations will be given to standing committees who advise on NAEP matters such as the Education and Information Advisory Committee (EIAC) of the CCSSO, and the NAEP NETWORK. Their counsel will be sought on matters of substance as the work of the Board progresses. The EIAC and other similar constituencies may also be invited to send a representative to all standing technical advisory committees of the Board's contractor(s) which deal with the level-setting process.

The Board will also seek advice from the technical community throughout the level-setting process. Efforts will be made to ensure that presentations are made regularly to such groups as the American Educational Research Association (AERA), the National Council for Measurement in Education (NCME), and the professional groups in the content areas such as the International Reading Association (IRA), the National Science Teachers Association (NSTA), and other similar organizations. The Board will seek to engage technical groups available to them, including the Technical Review Panel, the National Academy of Education, their own contractor(s), and NCES and its contractor(s), in constructive research studies focused on providing information on the technical aspects of NAEP related to level-setting (e.g., scaling, weighting, mapping ratings to the scale, etc.)

### **Validity and Reliability Evidence**

The Board will examine and consider all evidence of reliability and validity available. These data would include, but need not be limited to, procedural evidence such as the selection and training of judges and the materials and methods used in the process, reliability evidence such as intra-judge and inter-judge consistency data, and finally, internal and external validity data. Such data will help to inform *the Board's policy decision as they set the levels*.

Procedural evidence, while informative, is not necessarily sufficient evidence for demonstrating the validity of the levels. Therefore, the conduct of the achievement level-setting process shall be implemented so that a series of both internal and external validation studies shall be conducted simultaneously. To the extent possible, in order to realize maximum efficiencies in the use of resources, validation studies shall be included in the design of the level-setting data collection activities. Such studies may include, but shall not be limited to, convergent and divergent validation efforts, for example, conducting alternate standard-setting methods or conducting cross-validation level-setting activities, as well as exploring alternate methods for refining and expanding the preliminary achievement levels definitions, and empirically examining various technical decision rules used throughout the process.

As part of the validation task, additional evidence as to the suitability and appropriateness of identifying the subject area content of the recommended achievement levels ranges and cut-scores will be gathered. This evidence may include, but need not be limited to, data resulting from behaviorally anchoring the ranges and/or cut-scores, or data resulting from some other alternative procedures that employ a more global approach other than the item content of the particular assessment. The results of these studies will provide a clear indication of what students know and can do at the levels.

The results from these validation efforts shall be made available to the Board in a timely manner so that the Board has access to as much validation data as possible as it considers the recommendations regarding the final levels. Kane (1993) suggests that an “interpretive argument would specify the network of inferences leading from the score to the conclusions drawn about examinees and the decisions made about examinees, as well as the assumptions that support these inferences.” An interpretive argument which articulates the rationale for interpreting the levels shall accompany the presentation of proposed levels to the Board.

Again, to maximize the efficient use of resources and to minimize duplication of effort, it is highly desirable for contractors to coordinate the design of such studies with other agencies responsible for evaluating the level-setting activities.

## **Guideline 5**

The achievement levels shall be the initial and primary means of reporting the results of the National Assessment of Educational Progress at both the national and state levels.

## **Rationale**

In an effort to improve the form and use of NAEP the Board seeks to make the results of NAEP more accessible and understandable to the general public and to policy makers. The Board also supports the movement from norms-based assessments to standards-based assessments. Reporting the results of NAEP using the achievement levels accomplishes these ends to a greater degree than heretofore possible.

## Practices and Procedures

### Reporting What Students Know and Can Do

The purpose of most NAEP reports, but particularly those published under the auspices of the National Center for Education Statistics, is to report to the American public and others on the performance of students—that is, to report on *what students know and can do*. The purpose of the achievement levels is to identify for the American public what students *should know and should be able to do*, and to report the actual performance of students in relation to the achievement levels. Therefore, NAEP reports incorporate elements of both of these aspects of performance.

Clarity of interpretation of the NAEP data can be achieved by ensuring that the descriptions of performance for the levels and the exemplar exercises reflect what the empirical data show for a given assessment. This may be achieved by the modified procedures of *scale anchoring*<sup>1</sup> or by new procedures developed specifically for the purposes of providing elements of the content of the frameworks in the reporting mechanisms.

### Reporting Student Performance

In describing student performance using the levels, terms such as *students performing at the Basic level* or *students performing at the Proficient level* are preferred over *Basic students* or *Proficient students*. The former implies that students have mastery of particular content represented by the levels, while the latter implies an inherent characteristic of individual students.

In reporting the results of NAEP, the application of the levels of *Basic*, *Proficient*, and *Advanced* applies to the three regions of the NAEP scale generated when the appropriate cut scores are mapped to the scale. However, three cut scores yield, in fact, four regions. The region referenced by content which falls below the *Basic* cut score will be identified by descriptors that are not value-laden.

### Interpreting Student Performance

When interpreting student performance using the levels, one must diligently avoid over interpretations. For example, each of the NAEP subject areas are scaled independently of each other, even though each scale uses the same metric, i.e., scores ranging from 0 to 500. Because the metrics are identical, it does not follow that comparisons can be made across subjects. For example, a *Proficient* cut score of 235 in reading should not be interpreted to have the same meaning as a *Proficient* cut score of 235 in U.S. history. Neither should unwarranted comparisons be made in the same subject area from one assessment year to the next, unless the data for the two years have been equated and we have reason to believe that the scale itself has not changed from time 1 to time 2.

## **Guideline 6**

The level-setting process shall be managed in a technically sound, efficient, cost-effective manner, and shall be completed in a timely fashion.

## **Rationale**

Since a contractor(s) is conducting technical advisory and assistance work for the Board, it is critical that such work be performed to meet high quality standards, including efficiency, cost-effectiveness, timeliness, and adherence to sound measurement practices. *However, in the final analysis, it is the Governing Board that makes the policy decision regarding the levels, not the contractor.*

## **Practices and Procedures**

The contractor(s) shall prepare a fully detailed Planning Document at the onset of the level-setting work. This document will guide the progress of the work, serve as a monitor, and be the basis for staff and Board supervision. The Planning Document will outline milestone events in the process, provide a chronology of tasks and subtasks, as well as a monthly chronology of all activities across all tasks, and detail all draft and final documents that will be produced, the audience for such reports, and the number of copies to be provided by the contractor.

Procedures adopted by a contractor(s) to carry out the level-setting process must encourage and support national involvement by the relevant and required publics. Such meetings will also be conducted in a physical environment which is conducive to work and planning. To the extent possible, current technology shall be used in all areas of the level-setting process to increase efficiency and to reduce error.

The contractor(s) shall work closely and in a professional manner with the NAEP operations contractor in striving to fulfill the requirements of the level-setting process by (1) making all requests for information and data in a timely manner, (2) providing all requested information and data in a timely manner, (3) adhering to all predetermined deadlines so as not to impede the work of the operations contractor, and (4) advising the operations contractor of all unusual findings in the data so that a concerted effort can be mounted to resolve the problem or issue at hand.

The contractor(s) shall develop the initial level-setting design adhering to sound measurement principles and ensure that the various components of the design (e.g., selection of judges) are congruent with current standard-setting research. In the implementation of such designs, they shall employ state-of-the-art training strategies and measurement practices.

The contractor(s) shall produce documents in a timely manner and make oral presentations upon request. Presentations may include, but need not be limited to, the Board's quarterly meetings, relevant Board committees, and professional and lay groups.

## References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: APA.
- Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement (2nd ed., pp. 508-600)*. Washington, DC: American Council on Education.
- Jaeger, R.M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10*, 3-6, 10, 14.
- Kane, M. (1993). The validity of performance standards. Unpublished manuscript.
- National Academy of Education (1992). *Assessing student achievement in the states. The first report of the National Academy of Education panel on evaluation of the NAEP trial state assessment: 1990 trial state assessment*. Stanford, CA: Author.
- National Assessment of Educational Progress Authorization Act, (P.L. 110-279).
- Reid, J.B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice, 10*, 11-14.

## Endnotes

1. The traditional scale anchoring procedures anchored at the 200, 250, 300 350 points of the scale ( $\pm 12.5$  points), using a  $p = .65$ , and a discrimination of  $.30$  with the next lower level. The modified anchoring procedures (tried in reading for 1992) anchored at the achievement levels cut scores ( $\pm 12.5$ ), using a  $p = .65$ , and no discrimination criterion.