Alternative Approaches to Setting Performance Standards for the National Assessment of Educational Progress (NAEP)

Albert E. Beaton
Boston College

Robert L. Linn
University of Colorado at Boulder

George W. Bohrnstedt
American Institutes for Research

February 2012
Commissioned by the NAEP Validity Studies (NVS) Panel

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

This report was prepared for the National Center for Education Statistics under Contract No. ED-04-CO-0025/0012 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Albert E. Beaton
Boston College

Peter Behuniak
University of Connecticut

George W. Bohrnstedt
American Institutes for Research

James R. Chromy
Research Triangle Institute

Phil Daro
University of California, Berkeley

Lizanne DeStefano
University of Illinois

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Robert Linn
University of Colorado at Boulder

Ina V.S. Mullis
Boston College

Scott Norton
Louisiana Department of Education

Gary Phillips
American Institutes for Research

Lorrie Shepard
University of Colorado at Boulder

David Thissen
University of North Carolina, Chapel Hill

Karen Wixson
University of North Carolina, Greensboro

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Janis Brown
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
Phone: 650/ 843-8100
Fax: 650/ 843-8200

CONTENTS

The National Assessment Governing Board introduced the concept of achievement levels in reporting the 1990 NAEP assessment of mathematics. The judgmental methods used in producing the achievement levels have been both severely criticized and robustly defended. Even after 20 years of improvements, these methods occasionally give "unreasonable" results, as was the case with the 2009 science assessment, leading the controversy over them to arise again. Defenders of the achievement levels would argue that they have been of key importance in setting much needed standards for performance that were lacking with the previously used NAEP anchor points. They would further argue that the achievement levels have been especially useful for looking at the changes in Black-White and Hispanic-White achievement gaps, as well as for establishing a common metric against which the proficiency levels set by states can be compared.[1] It is also clear that different judgmental methods can yield different results. Further, a given judgmental method can yield different results when used in different contexts, when used in different content areas, or when different judges set the standards.

The arguments for and against the current achievement levels both have merit. While the continuing concerns about the current approach to setting achievement levels offer one reason to consider whether there are better approaches to setting achievement levels, there are at least three other reasons as well. First, the Governing Board is in the process of attempting to set a new type of performance standards at grade 12—called preparedness levels. The goal is to determine whether cut points can be set for preparedness "to qualify for entry-level college courses leading to a degree or job training without the need for remediation" (Fields, 2011). If done successfully, it can be argued that these levels might replace the current grade 12 achievement levels, given their external validity. Second, the move towards Common Core Standards by the states almost certainly will have implications for the NAEP frameworks and hence for the assessments and achievement levels as well. Finally, because issues about the validity of the achievement levels remain unresolved, the most recent reauthorization of ESEA requires that they be used on a trial basis until the Commissioner of the National Center for Education Statistics (NCES) determines they are "reasonable, valid, and informative to the public." In the meantime, their trial status is to be noted in each NCES report where they are used (No Child Left Behind Act of 2001, P.L. 107-110, 115 Stat. 1425 [2002]). The Commissioner has recently noted that he is interested in addressing the matter of whether the trial status of the achievement levels should be removed.

For all of these reasons, it is a propitious time to consider how achievement levels on NAEP are set. In this paper we first examine the background of how the current achievement levels were set. We then explore three possible alternatives to setting achievement levels. The first alternative relies on making the achievement levels predictive—as is being done in setting levels for grade 12 college and work preparedness. The second alternative is benchmarking the achievement levels against international standards, and the third involves using percentiles to set base-year

[1] Importantly, however, the size of the gaps or changes in performance from one assessment year to another depends on where the cut points are set (Holland, 2002).

achievement levels. Hybrid approaches that combine features of these methods are also discussed.

Background

From its earliest days, NAEP has focused on what American students know and can do. Early NAEP avoided creating any student scale scores, even simple number-correct scales. The basic idea was to publish item-level results from a small sample of a large pool of assessment items. The results for an individual item would be presented with one or more interpretations. However, the lack of generalization over the many assessed items limited the value of NAEP reports for policymakers and the general public. Soon, some summarization was introduced by averaging statistics such as the percentage correct over many items, but no effort was made to produce student scores. In fact, such scores were studiously avoided.

"New Design for a New Era" (Messick, Beaton, & Lord, 1983), published when ETS took over the design of NAEP, focused on Item Response Theory (IRT) scaling as a way of summarizing student performance data. The first IRT analysis in NAEP was done in the 1983–84 reading assessment. The new reading scale was a vertical scale covering 9-, 13-, and 17-year-old students. It was designed to have an initial overall mean of 250 and a standard deviation of 50. Progress in student performance could be viewed over time by computing averages of the scale scores and/or the percentage of students who performed at or above various score points.

Developing a meaning for the scale was important if NAEP was to communicate its findings effectively to policymakers and the general public. To do so, NAEP developed scale anchoring (Beaton & Allen, 1992). This procedure selected several scale points for describing student performance: 150, 200, 250, 300, and 350. These "anchor" points were arbitrary but not capricious in that they covered the range of student performance reasonably well. Each anchor point (except 150) was described by saying what students at that point knew and could do that most students at the next lower anchor point could not. Illustrative items were shown for each anchor point. Such scale anchoring also has been used in the Trends in Mathematics and Science Study (TIMSS) and is now used for describing proficiency in the national assessments of several countries around the world.

The Governing Board first set judgmental performance standards, referred to as achievement levels, on the 1990 mathematics assessment. The achievement levels became the subject of considerable controversy almost immediately. They were strongly criticized in an evaluation report commissioned (and later rejected) by the Governing Board (Stufflebeam, Jaeger, & Scriven, 1991). The controversy that began in 1991 has not abated over the two decades that the achievement levels have been in existence and used in the reporting of NAEP results. The most recent evaluation of NAEP (Buckendahl et al., 2009) argued that the consumers of NAEP data appreciate the use of achievement levels in reporting, but went on to acknowledge that the setting of achievement levels remains controversial—noting that the process of setting achievement levels had been criticized in earlier evaluations (e.g., Shepard, Glaser, Linn, & Bohrnstedt, 1993; U.S. General Accounting Office, 1993; Pellegrino, Jones, & Mitchell, 1999) but defended by others (e.g., Cizek, 1993; Hambleton et al., 2000; Loomis & Bourque, 2001; Bourque, 2004).

It is worth noting that the insistence on a criterion-referenced approach with fixed cut points and the reliance on the primary reporting of results in terms of the percentage of students scoring above selected cut points (e.g., the percentage of students that is proficient or above) have substantial limitations. In particular, the percent-above-cut statistics can paint quite different pictures of trends or in the gaps in achievement between two subgroups (Holland, 2002). These limitations apply not only to the standard-setting methods that have been used in the past, but to all of the methods discussed in this paper. Thus, any approach that leads to reliance on percent-above-cut statistics has drawbacks when compared to more traditional approaches using means and effect-size statistics.

Over the years, the Governing Board has attempted to improve the process of setting achievement levels. For example, the Governing Board moved from a modified version of the Angoff method—the method used in the initial rounds of standard setting that received a great deal of criticism—to a modified bookmark method known as the Mapmark method (Yin & Sconing, 2008). However, the NCES disclaimer remains in effect. The tentative status of the achievement levels is particularly troubling because of the use that has been made of the NAEP achievement levels as the primary benchmark for evaluating the performance standards that states are required to set on their assessments under the *No Child Left Behind* (NCLB) Act of 2001. (See Bandeira de Mello, Blankenship, and McLaughlin, 2011, for the most recent report benchmarking state performance standards against NAEP.) It is also troubling that—for various subject areas and grades—the standard-setting process has sometimes led to recommendations of cut scores for the advanced levels that are so high that no one or virtually no one scores at the advanced level.

Suggestions have been made for improving the standard setting on NAEP (see, for example, Pelllegrino, Jones, and Mitchell, 1999, chapter 5), but they have either not been followed or have not led to marked improvements. The revised methods still sometimes lead to “unreasonable” results such as no students scoring in the advanced category. Moreover, they have not overcome the problem that the standards set in different subjects yield quite different percentages of students scoring at or above the three cuts that define the basic, proficient, and advanced levels of performance. This is an important limitation because it gives the impression that students are performing at a lower level in one content area than another when the difference may be due to differences in the stringency of the standards that are set in the different content areas.

As suggested in the introductory comments, because of the Governing Board’s current effort to set preparedness levels at grade 12, the potential for other changes to NAEP that the Governing Board may deem appropriate as the nation moves towards Common Core Standards, and the Commissioner’s interest in whether the trial status of the current levels should be lifted, this is a particularly good time to consider whether and how achievement levels might be changed. *This paper does not pretend to solve the problem of how new achievement levels should be set. Instead, its purpose is to initiate thinking about the challenge and opportunity that the current context presents for rethinking the process.*

Predictive Achievement Levels

The current achievement levels are set judgmentally through a process in which participants decide—without any reference to prediction—what students should know and should be able to do at given points in their school careers, namely grades 4, 8, and 12. The quality of the resulting achievement levels is evaluated by a set of studies to ensure they are internally valid. External validity is ignored. By contrast, a predictive approach places a high priority on the external validity of the cut points against predetermined criteria. For grade 4 mathematics, for example, performance in grade 5 mathematics would seem the most logical criterion. Similarly, for grade 8 reading, performance in grade 9 English Language Arts would seem to be the most logical criterion. The predictive approach may be most appropriate for grade 12, where, as was mentioned above, the Governing Board has an expressed interest in setting preparedness standards for college-level work and job training programs, but research could be done to determine the predictive validity of the achievement levels at grades 4 and 8 as well.

Unfortunately, there have been few studies to date to examine the predictive validity of NAEP, and especially the predictive validity of the achievement levels. The few studies that have been done have focused on the grade 12 or age 17 samples. The 1983–84 NAEP Technical Report (p. 536) presented the correlations between the NAEP reading plausible values and the PSAT Verbal and Quantitative scores for students who had taken both exams. The correlation was .67 with the PSAT Verbal scores and .57 with the PSAT Quantitative scores. This is important validity evidence. Analysis of correlations between the grade 12 NAEP plausible values and SAT scores is currently in progress at ETS as part of the Governing Board's research program examining the feasibility of setting preparedness levels, and preliminary results look promising. Future studies might also use NAEP results to predict college grades, which would provide a means of determining the score at which students have, say, a 60 percent chance of earning a grade of C or better.

NCES is examining how the NAEP achievement levels relate to the SAT and ACT scales for college success, which have had many validity studies indicating how well entering students with different score levels do in college. With linkages to SAT and ACT, the NAEP 12th grade achievement or preparedness levels could be used to estimate how well graduating high school seniors could be expected to perform in college. The Governing Board also has a study underway that takes advantage of longitudinal data collected on students in the state of Florida using the overlap sample of students who also took grade 12 NAEP. The intention is to examine the relationship between student achievement on NAEP and students' college placement score results (as well as with employment data, including salaries).

An important issue for all of these studies is getting agreement on the criterion for college success. The Governing Board defines college preparedness as being ready to take credit-bearing college courses without remediation. Presumably this means earning at least passing grades. Others might suggest that the criterion should be higher—getting a B- or better with a 50 percent probability, or a C+ or better with a 75 percent probability, for example. We address this issue further below.

While the Governing Board has not indicated what will become of the current grade 12 achievement levels if their attempts to create cut points for college and work preparedness are successful, such a success would call into question the need for the current achievement levels. It is for this reason that we suggest it is a good time to examine whether achievement levels at *all three grades* should be based on their ability to predict meaningful criteria. However, while the grade 12 studies that the Governing Board is carrying out will provide useful data on the external validity of the grade 12 achievement levels, to our knowledge there have been no studies to date to show whether the grade 4 achievement levels predict grade 5 performance or whether grade 8 achievement levels predict high school performance.

Interesting validity evidence would be available if adjacent grades were assessed—the 7th and 8th grades, say. This would allow estimation of average growth between the grades. The states or jurisdictions would be compared as usual, but now there would be an additional yardstick: the average growth between grades. The average growth could be used as a measure of size for state comparisons as well as for providing new information about improvements in student performance. The adjacent-year assessments could also provide a more fine-grained look at changes in the percentages of students at different achievement levels, as well as other factors such as the proportions of students being promoted or held back, or the proportions moving into advanced courses such as geometry or algebra.

Adjacent-grade assessment would be expensive, but perhaps not as expensive as one might think. The assessment booklets would be exactly the same for both grades. The within-school sample size would need to increase, but the sample of schools would be the same as for the usual NAEP assessments and so the assessors would need to go there anyway. Adjacent grades that were not in the same school buildings would add manageable difficulties.

The data from adjacent-grade sampling would result in substantial gains in NAEP interpretability. At present, the scales for main NAEP are not really vertical scales, although they tempt the user to treat them as such (Haertel, 1991; Thissen, 2012). Testing grades 4 and 5, 7 and 8, and 11 and 12 would give estimates of growth at three levels. If the approach were expanded to testing all grades between 4 and 12, a truly vertical scale could be developed. However, in order to accomplish this, it would be necessary to prepare additional grade-appropriate items.

In general, when setting predictive achievement levels, one needs to "begin with the end in mind." That is, be clear about what criterion or criteria one wants to predict (e.g., a probability of 50 percent of completing the first year of college with a B average or higher) and then find the NAEP cut point associated with the criterion. One would want to do the same when setting achievement levels for grade 8 and grade 4. For example, at grade 8 mathematics "proficient" might be the cut point associated with a B or better in specific grade 9 mathematics courses (e.g., Algebra I). Again, specific criteria should be specified to be predicted for a given subject area before arbitrarily deciding that there should be three achievement levels. What would be the three criteria one is trying to predict for grade 5 mathematics performance, for example? Would the three criteria be earning an A, B, or C in 5th grade mathematics? Or is there really only a single criterion such as earning a grade of B or better? If the latter, then a single cut point makes sense.

The predictive approach to setting performance standards has some potential advantages over judgmental approaches, but it also has some disadvantages. Some potential advantages and disadvantages, or pros and cons, are noted below.

Two possible advantages or pros are as follows:

1. The predictive approach provides a link to criteria that people care about and for which they have some experience-based understanding.
2. The predictive approach uses an external frame of reference that may also be useful in understanding the validity of interpretations of NAEP scores.

These pros need to be considered along with the possible disadvantages or cons of the predictive approach.

Some cons of the predictive approach are as follows:

1. It requires judgments of the course grade that is considered good enough or acceptable.
2. It would be expensive to get the needed data.
3. It may be less applicable at grades 4 and 8 than at grade 12.
4. It may give the possibly misleading impression that students are performing less well in one content area than another, but at least this would be grounded in differences in student performance in different subject areas.
5. It requires that the cognitive and content requirements associated with meeting the joint AERA/APA/NCME criterion for predictive validity align with the cognitive requirements and the content of NAEP. (American

Educational Research Association, American Psychological Association, & the National Council for Measurement in Education, 1999)

Benchmarking to International Standards

An alternative criterion-based approach would be based on linking NAEP to TIMSS, the Progress in International Reading Literacy Study (PIRLS), and the Programme for International Student Assessment (PISA). There is great interest in knowing how our students' performance compares with that of students from other countries. This is concurrent rather than predictive validity, but it, too, is a form of external validity. Linking could be used to map the cut scores used by TIMSS, PIRLS, or PISA to the NAEP scale. Item anchoring could then be used to provide substantive meaning to the cut scores. Phillips (2007) reviews several previous efforts to link the NAEP and TIMSS scales and provides a linking that projects the NAEP scale points onto the TIMSS scale. His approach has been used successfully to map the NAEP achievement levels onto the TIMSS scale and has also been used in a few states to help guide the standard-setting process on the assessments used in those states.

If Phillips' approach was used to link NAEP to either TIMSS or PISA, the linked cut points used for reporting results on the international assessments could be used to identify corresponding points on the NAEP scale. In this way, NAEP results could be reported in terms of the percentage of students who score at or above the various international benchmarks on TIMSS, PIRLS, or PISA.

Of course, there is nothing that prohibits using various criteria. That is, one can have cut points for proficient performance when predicting 5th grade reading performance on reading *and* one can have proficient performance for linking to grade 4 PIRLS reading, for example. The main disadvantage of the international benchmarking approach is that a more limited range of subjects and grade levels are assessed by the international assessments than are assessed by NAEP.

Some of the pros of the international benchmark approach are as follows:

1. It provides a frame of reference that is of interest to policymakers and the public.
2. It is less expensive than the predictive approach.

Some of the cons of the international benchmark approach are that

1. It cannot be used in grades and subjects not tested in international assessments.
2. It depends on standards set in international assessments that may have limited validity and may be hard to interpret.
3. Links to international assessments may be weak either because of the lack of common test-takers or common items.

Baseline Normative Standards

A totally different approach from the predictive and concurrent international benchmark approaches described above is the use of normative standards. Normative comparisons are anathema to many proponents of criterion-referenced or standards-based reporting of assessment results. However, normative comparisons over time are fundamental to measuring progress on NAEP or any other assessment used to evaluate trends in achievement. Positive progress has been made when achievement in one year is higher than it was in a previous year. That is, the results in a previous year provide the normative basis for judging performance in the current year. This approach is common practice in interpreting long-term trend NAEP results, for example.

Thus, the normative results in a baseline year—for example, achievement on the NAEP mathematics assessment in 2000—could provide the desired performance standards for reporting achievement in 2011. In keeping with current practice, three cut scores could be established to correspond to selected percentile ranks (e.g., 95th, 50th, and 25th percentiles) in the baseline year. Labels could be given to the four score regions defined by the three cut scores (e.g., advanced, proficient, basic, and below basic), or, to avoid confusion with the achievement levels, the four score regions could be labeled exceeds standard, meets standard, meets minimal expectations, and unsatisfactory, or simply levels 4, 3, 2, and 1.

Of course, gains can also be seen using the current achievement levels, but standards set using percentile points in a base year have the advantage that the starting percentages are easy to remember. In addition, the anchoring process described below would produce descriptions of what students at various levels actually knew and were able to do.

Few would argue that predictive, criterion-related achievement levels are not more defensible than the current achievement levels, which, while internally valid, lack any criteria for judging how good or bad they are. But both types of achievement levels are expensive to set. One approach would be to approach the setting of predictive achievement levels incrementally. Assuming that the setting of preparedness levels at grade 12 is successful, the next step would be to do studies to make the grade 8 achievement levels predictive. If even this is viewed as just too expensive, however, the baseline normative approach is an alternative for which the cost is virtually negligible. Before applying the baseline normative approach, however, it is important that some research be done on the details of the process, including what percentiles to select as cut points.

Some of the pros of the baseline normative approach are as follows:

1. Normative comparisons are familiar to the public and used not only in testing but in other day-to-day contexts such a pediatrician's discussion of a child's height or weight with parents.
2. The approach would be inexpensive to implement.

3. It would, by definition, yield comparable percentages of student in each performance level across different content areas in the base year.
4. Changes in student achievement would be evident from the results each year without the need to know the percentage of students who were proficient or above in the year the proficient standard was set using a judgmental method.

Some cons of the baseline normative approach are that

1. Normative comparisons are eschewed by many (in some cases, unfairly).
2. The approach does not answer the question, "how good is good enough?"
3. Choosing percentiles can be arbitrary, and different percentiles will lead to different trends and gaps in the performance of subgroups.

Hybrid Methods

In addition to the current judgmental standard-setting method and the three methods described above, there are, course, hybrid methods that could be used to capitalize on the advantages of the different methods. One hybrid method combines the benchmark method with a content-based judgmental approach. Phillips (2011) has proposed such an approach that he has called "the benchmark method of standard setting." This method is grounded in content standards or content frameworks, but is not solely based on content considerations. Before reviewing items to set judgmental standards, judges are given information regarding the percentage of students scoring above various international performance standards, based on a link of the subject assessment with an international assessment.

The baseline normative method could be easily combined with either the predictive or benchmark methods described above. An approach that combined the predictive with the baseline normative methods would have many of the advantages of the two methods and would eliminate some of the disadvantages of each. The predictive method would be used to set 12th grade performance standards using grades in first year college courses leading to a degree. This would be done in just one content area—either mathematics or reading. The percentage of students performing above the identified cut points determined for the selected content area would be used to establish the baseline normative standards that would then be applied to set the standards for other grade 12 content areas and for all content areas at grades 4 and 8.

This hybrid approach would have the advantage that the performance standards would be grounded in the external reference of college performance that is familiar to policy makers and the public and that they care about. It would also have the advantage that standards at different grades and different subject areas would be comparable and therefore would not give the possibly misleading impression that students are performing better in one subject area than another.

The international benchmark and baseline normative methods could also be combined in a parallel manner. We believe, however, that the combination of the predictive and normative approaches has advantages over all of the other approaches considered in this paper. Therefore, we prefer the combined predictive and baseline approach and recommend that it be given serious consideration.

An Endnote on Better Describing What Students at a Given Level Know and Can Do

Standard setting on NAEP assessments implies setting score cut-points that separate the various achievement levels. Describing what students at various achievement levels know and can do is fundamental to understanding what is required to meet the standards. This section proposes a method for interpreting and describing such achievement levels—*the level description method*.

To provide a context, when the Governing Board develops a new framework, one of the products is a preliminary description of what students should know and be able to do if they perform at the basic, proficient, or advanced levels. Subsequently, when committees are put together to set achievement levels in a given content area,

> 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. (National Assessment Governing Board, 1995, p. 7)

That is, the method produces an expanded definition of what *students* at grades 4, 8, and 12 should know and be able to do in a given content area (e.g., science) for the basic, proficient, and advanced levels, along with some exemplar items. By way of contrast, the level description method described here provides information about where each *item* performs and uses that information to provide descriptions of actual performance at each level. In addition, the method provides information about what students at the below basic level can do. The current achievement level descriptions can only report on what below basic students *cannot* do. Given the interest from both policymakers and practitioners about the significant percentages of students at the below basic level, this alone would seem to be a compelling reason to consider the level description method to which we now turn.

The level description method is similar to the scale anchoring method, but differs in a substantial way. The level description method looks for items that discriminate between students at different levels, not at particular scale points. Both methods assume that a large sample of students has been assessed and that the students' scale scores (or plausible values) and item responses are available for analysis. We also assume here that initial cut points are available and may be adjusted if they are found to be unsatisfactory.

The first step in interpretation is finding assessment items that discriminate between adjacent achievement regions. Discriminating items have the property that a large percentage of students at an achievement level can respond correctly to the item whereas most students at the lower levels cannot. Operationally, we proposed defining a "large percentage" as over 80 percent and "most students at the lower levels" as less than 50 percent. The 80 percent and 50 percent values are quite stringent, however, and may need to be modified.[2]

Assume that there are three cut points separating four levels as in the present NAEP achievement level setting process. We will label them 1, 2, 3, and 4, with a 5th category for items that do not discriminate among the performance levels. For the present achievement levels, these would be named below basic, basic, proficient, and advanced (where 1 = below basic, 2 = basic, and so on).

Which items discriminate among the levels? The assessment items could be sorted into the following groups:

1. Below basic items, where 80 percent of the students who scored below basic answered correctly.
2. Basic items, where 80 percent of the students who scored in the basic range answered the items correctly but fewer than 50 percent of the "below basic" students answered them correctly.
3. Proficient items, where 80 percent of the students who scored in the proficient range answered the items correctly but fewer than 50 percent of the "below basic" and "basic" students answered them correctly
4. Advanced items, where 80 percent of the students who scored in the advanced range answered the items correctly but fewer than 50 percent of the "below basic," "basic," and "proficient" students answered them correctly
5. Non-discriminating items, which do not discriminate among the achievement levels.

The item groups associated with performance levels would then be reviewed and interpreted.[3] Both concise and detailed descriptions of the differences in performance among the student groups would be developed.

The non-discriminating items may also be of interest. They may suggest, for example, that the cut points are too close and that separating the cut points could result in more distinct discrimination patterns and clearer interpretations.

A NAEP report would include not only the percentage of students at each level but also a description of what those students actually knew and could do, based on the

[2] These values should be thought of as only starting points. Other values may do a better job of identifying a useful number of discriminating items and make better sense.

[3] For partial credit response items, each response category would be treated as a separate item, just as is done now.

actual items in that assessment. The report would also present selected discriminating items to describe the differences among the levels. Changes in student performance on individual published discriminating items could be reported as well.

Knowing what students actually know and can do can help in the establishment of aspirational achievement levels. Inspecting the student results may well suggest higher cut points that can reasonably be attained in future assessments. In addition, the review of the discriminating items may suggest changes in curricula or teaching strategies to improve student performance; for example, if certain geometry items in a mathematics assessment discriminate between the most advanced students and the others, then curriculum in this area may be modified and taught so that more students attain the advanced level.

Summary

The judgmental approaches used by the Governing Board to set NAEP achievement levels have been the subject of a great deal of controversy and have sometimes led to results that were considered “unreasonable.” The NCES requirement that the achievement levels be used on a trial basis and be interpreted with caution continues to be in effect. For these reasons and because substantial changes in NAEP are likely to occur in the next few years due to the introduction of the 12th-grade preparedness levels and the widespread adoption of the Common Core Standards by states, this seems a propitious time to consider alternative approaches to setting performance standards.

Three alternative approaches to the usual judgmental approaches were described in this paper. These are the predictive, the international benchmark, and the baseline normative approaches. The predictive approach would identify score levels that correspond to defined subsequent outcomes such as having a probability of getting Cs or better in credit-bearing college courses leading to a degree. The international benchmark approach would use international assessments such as PISA, TIMSS, or PIRLS to establish achievement levels by linking NAEP to the international assessment. The baseline normative approach would use norms in a baseline year and choose percentile points such as 95, 75, and 50 to define achievement levels that would be used in future assessments. Some pros and cons of each of the three alternative approaches were presented.

Some hybrid approaches that combine features of different methods were also described. One hybrid approach has a number of advantages: It starts by using the predictive method to set cut points for a single content area at grade 12, using first-year college grades as the criterion. As the next step, the percentiles corresponding to these predictive cut points are used to set standards in other content areas and grades. We believe this combination of the predictive and baseline normative methods deserves serious consideration.

Finally, a level description method was described that could be used with a judgmental approach, or any of the other approaches discussed, to give substantive meaning to the levels.

References

American Educational Research Association, American Psychological Association, & the National Council for Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. (2011). *Mapping state proficiency standards onto NAEP scales: Variation and change in state standards for reading and mathematics (2005–2009).* (NCES 2010-458). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Beaton, A. E., & Allen, N. (1992). Interpreting scales through scale anchoring. *Journal of Education Statistics*, *17*(2).

Bourque, M. L. (2004). A history of the National Assessment Governing Board. In L. V. Jones and I. Olkin (Eds.), *The nation's report card: Evolution and perspectives* (pp. 201–231). Bloomington, IN: Phi Delta Kappa Education Foundation.

Buckendahl, C. W., Davis, S. L., Plake, B. S., Sireci, S. G., Hambleton, R. K., Zenisky, A. L., & Wells, C. S. (2009). *Evaluation of the National Assessment of Educational Progress: Final report.* Buros Institute for Assessment Consultation and Outreach, Buros Center for Testing, University of Nebraska–Lincoln, & Center for Educational Assessment, University of Massachusetts Amherst.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, *30*, 93–106.

Fields, R. (2011). *NAEP 12th grade preparedness: The Governing Board program of research.* [PowerPoint presentation]. Washington, DC: National Assessment Governing Board.

Haertel, E. H. (1991). *Report on TRP analyses of issues concerning within-age versus cross-age scales for the National Assessment of Educational Progress.* Washington, DC: National Center for Educational Statistics. Retrieved from http://www.eric.ed.gov:80/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED404367&ERICExtSearch_SearchType_0=no&accno=ED404367

Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., … Zwick, R. (2000). A response to "setting reasonable and useful performance standards" in the National Academy of Sciences' "Grading the Nation's Report Card." *Educational Measurement: Issues and Practice*, *19*(2), 5–14.

Holland, P. W. (2002). Two measures of change in gaps between CDFs of test score distributions. *Journal of Educational and Behavioral Statistics*, *27*(1), 3–17.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods and perspectives* (pp. 175–217). Mahwah, NJ: Erlbaum.

Messick, S. J., Beaton, A. E., & Lord, F. M., (1983) *A new design for a new era.* Princeton, NJ: Educational Testing Service.

National Assessment Governing Board. (1995). *Developing student performance levels for the national assessment of educational progress: Policy statement.* Washington, DC: National Assessment Governing Board.

No Child Left Behind Act of 2001. Public Law 107-110.

Phillips, G. W. (2007). *Expressing international educational achievement in terms of U.S. performance standards: Linking NAEP achievement to TIMSS*. Washington, DC: American Institutes for Research.

Phillips, G. W. (2001). The benchmark method of standard setting. In G. J. Cizek (Ed.), *Setting Performance Standards* (2nd edition). New York: Routledge.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.). (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.

Shepard, L., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement: A report of the National Academy of Education panel on the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels.* Stanford, CA: National Academy of Education.

Stufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1991). *Summative evaluation of the National Assessment Governing Board's Inaugural 1990–91 effort to set achievement levels on the National Assessment of Educational Progress.* Prepared for the National Assessment Governing Board, August 23.

Thissen, D. (2012). *Validity issues involved in cross-grade statements about NAEP results.* San Mateo, CA: NAEP Validity Studies Panel.

U.S. General Accounting Office. (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations*. GAO/PEMD-93-12. Washington, DC: U.S. General Accounting Office.

Yin, P., & Sconing, J. (2008). Estimating variability of cut scores for item rating and Mapmark procedures: A generalizability theory approach. *Educational and Psychological Measurement*, *68*, 25–41.