A Response to "Setting Reasonable and Useful Performance Standards" in the National Academy of Sciences' *Grading the Nation's Report Card*

Ronald K. Hambleton,
University of Massachusetts, Amherst

Robert L. Brennan,
University of Iowa

William Brown,
Brownstar

Barbara Dodd,
University of Texas, Austin

Robert A. Forsyth,
University of Iowa

William A. Mehrens,
Michigan State University

Jeff Nellhaus, *Massachusetts Department of Education*

Mark Reckase,
Michigan State University

Douglas Rindone,
Connecticut Department of Education

Wim J. van der Linden,
University of Twente, The Netherlands

Rebecca Zwick,
University of California, Santa Barbara

Our Position on the NAEP Performance Standards

Since 1990, the National Assessment Governing Board (NAGB) has been setting performance standards (called *achievement levels* by NAGB) on the National Assessment of Educational Progress (NAEP). Performance standards have been set on the 1990 and 1992 Mathematics, 1992 Reading, 1994 U.S. History, 1994 Geography, 1996 Science, 1998 Civics, and 1998 Writing assessments. The purposes of the performance standards are to (a) provide a frame of reference in which policymakers, educators, and the public can understand the NAEP test results; (b) provide more interpretive information about the meaning of NAEP test scores by defining three categories of performance: basic, proficient, and advanced; and (c) promote excellence in education. Standard-setting on NAEP has been controversial—initially, critics were opposed to both performance standards *and* the process used for setting them. Today, the performance standards in NAEP score reporting appear to be widely accepted by policymakers and the public, but the process for setting them remains controversial.

Our review of the evidence for the NAEP performance standards indicates there is support for using the current methods for setting performance standards on NAEP and other national and state assessments. The NAGB standard-setting methods themselves have been carefully and systematically developed and improved through numerous

Ronald K. Hambleton is Distinguished University Professor in the School of Education, University of Massachusetts 01003-4140. His specializations are large-scale assessment and item response theory.

Robert L. Brennan is E. F. Lindquist Professor of Measurement and Director of the Iowa Testing Programs at the University of Iowa, 334 Lindquist Center, Iowa City, IA 52242. His specializations are generalizability theory, equating, and large-scale assessment issues.

William J. Brown is President of Brownstar, Inc., an educational services company. His specializations are the administration of large-scale testing programs, evaluation, and standard setting.

Barbara Dodd is Professor of Education at the University of Texas, Austin. Her specialization is computer adaptive testing.

Robert A. Forsyth is Professor of Educational Measurement and Statistics at the University of Iowa, Iowa City 52242. His specializations are the development and use of achievement tests.

William A. Mehrens is Professor of Educational Measurement at Michigan State University, 462 Erickson Hall, East Lansing, MI 48824. His specializations are consequences of and legal issues in large-scale assessments.

Jeffrey M. Nellhaus is the Director of Student Assessment for the Massachusetts Department of Education, 350 Main St., Malden, MA 02148. His specialization is the implementation of large-scale, standards-based, student assessment programs.

Mark D. Reckase is Professor of Measurement and Quantitative Methods at Michigan State University, East Lansing, MI 48824. His specializations are multidimensional item response theory, computerized adaptive testing, and implementation of large-scale assessments.

Douglas A. Rindone is the Chief of the Bureau of Student Assessment and Research, Connecticut Department of Education, Hartford, CT 06145-2219. His specializations are in large-scale student assessments and program evaluation.

Wim J. van der Linden is Professor of Educational Measurement and Data Analysis, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. His specializations are item response theory and computerized testing.

Rebecca Zwick is Professor, Department of Education, University of California, Santa Barbara 93105-9490. Her specializations are applied statistics, psychometrics, and the assessment of test fairness.

research studies, pilot-tests, and implementations over eight years. Every aspect of the process has been researched and carefully implemented in the standard-setting methodology, from the selection of broadly representative groups of panelists and advanced preparation for the standard-setting meetings, to the development of performance descriptors, to the training of panelists and implementation of the methods in a 5-day process, to the careful evaluation and documentation of the process.

Our positive evaluation of the NAEP standard-setting methodology is clearly at odds with the very negative evaluation of this same methodology reported by the National Academy of Sciences (NAS) in Chapter 5 of *Grading the Nation's Report Card* (Pellegrino, Jones, & Mitchell, 1999). But, in our view, "Chapter 5: Setting Reasonable and Useful Performance Standards," presents a very one-sided and incomplete evaluation that is based largely on dated and second-hand evidence. The faulty conclusions and incomplete review of evidence to support the conclusions suggest to us that the authors of Chapter 5 (a) placed too much emphasis on dated material, (b) ignored the extensive body of research that supports the validity of the NAEP performance standards, and (c) were relatively uninformed about the challenges associated with the NAEP standard-setting process. In our judgment, Chapter 5 of the NAS Report does not conform to generally accepted scientific standards of objectivity in conducting research or evaluation. In sum, we are surprised, dismayed, and disappointed with the inadequate scholarship reflected in the review of NAGB's 8 years of research and development to set performance standards on NAEP. Policymakers, educators, and the public deserve a better review and evaluation than they received for their substantial investment of research funds.

Goals and Conclusions of the NAS Report

According to the authors of Chapter 5, their goals were the following (p. 163):

1. To provide an overview of NAEP's performance standards and the achievement-levels-setting process as it was conducted through 1996.
2. To summarize the major findings of previous evaluations and research efforts that have examined this process.
3. To present a detailed accounting and evaluation of the achievement-levels-setting process as it was applied to the 1996 NAEP Science Assessment.
4. To present the committee's conclusions and recommendations about standard-setting.

These authors arrived at the following summary conclusion: "Standards-based reporting is intended to be useful in communicating student results, but the current process for setting NAEP achievement levels is fundamentally flawed" (p. 162). We note, however, that the phrase "fundamentally flawed," used throughout the NAS Report as an overall evaluative conclusion, is never justified in any serious scientific sense. This mantra has certainly captured the attention of the public, the press, and the Congress. For example, Rothstein (1998) uses this phrase on page 73 of his work. As a second example of the widespread use of the phrase, consider Bracey's recent *Phi Delta Kappan* article, "The Ninth Bracey Report on the Condition of Public Education" (1999). Bracey uses this phrase on p. 155: Rhetoric should *not* be substituted for scientific inquiry, and continued repetition of an unwarranted conclusion does not enhance its validity.

Our Response to Chapter 5 of the NAS Report

The primary purpose of this article is to respond to the criticisms of the NAEP standard-setting methodology put forth in Chapter 5 of the NAS Report. Specifically, the objectives of this response are as follows:

1. Identify several points made in Chapter 5 with which we agree.
2. Offer several general criticisms of the arguments and interpretations of data offered in Chapter 5.
3. Address the specific criticisms regarding the NAEP standard-setting process noted in Chapter 5.
4. Consider the evolution and integrity of the achievement-levels-setting process that has been used with NAEP assessments.

Subsequent sections address the four objectives noted above.

Points of Agreement

There are a number of observations and comments in Chapter 5 with which we concur. The most important of these are discussed below.

We definitely agree that standards-based reporting "has the potential to be a significant improvement in communicating about student achievement to the public and to policy makers" (p. 171). Most state departments of education also agree, and currently use standards-based reporting in their own assessments. Furthermore, we strongly support the call for additional research to (a) help improve the use and interpretation of standards-based reporting, (b) assess the impact of standards-based reporting on curriculum development and instruction, and (c) continue to explore the usefulness of alternative standard-setting procedures. In fact, as noted later in this article, a considerable amount of research related to alternative standard-setting procedures has been undertaken by NAGB, and more is being planned.

We also support the recommendation in Chapter 5 that preliminary achievement levels descriptions "should guide the development of assessment items and exercises" (p. 177). In fact, in recent assessments (e.g., civics), it is our understanding that such descriptions have played an important role in the development of the assessment.

We concur with the authors of Chapter 5 that all standard-setting methods rely heavily on informed judgments and that "true achievement levels do not exist" (p. 173). We were surprised, however, to see the authors of Chapter 5 offer this as one of their conclusions. In fact, this view has been in the measurement literature for more than 20 years, and it is accepted by every knowledgeable measurement specialist known to us who works in the area of standard setting.

In addition, we support the conclusion that despite the judgmental nature of standard-setting, the process "whereby decisions are reached in achievement levels setting should be well documented and clearly communicated" (p. 173). In a later section of this article, we argue that considerable documentation exists for the standard-settings that have been undertaken for the NAEP assessments. The NAS evaluation reveals no evidence that suggests that available documentation for the 1996 Science achievement-levels-setting process has been considered.

General Criticisms

Before reacting to the more substantive issues in Chapter 5, it seems appropriate to make a few observations about the general style of argumentation and writing relative to the mission of the NAS. For example, one goal of the NAS is "to survey the broad possibilities of science, to formulate comprehensive projects of research, and to develop effective means of utilizing the scientific and technical resources of the country for dealing with such projects" (National Academy of Sciences, 1998, p. 18). Adherence to this goal should have resulted in a report based on such generally accepted scientific principles as a balanced and complete referencing of the literature, a careful consultation of primary sources of information, an explicit flow of argument based on clear premises, and a judicious evaluation without unwarranted prejudices. Instead, we believe Chapter 5 represents the results of an evaluation study in which each of these principles is ignored. This conclusion is substantiated by our observations below.

Standard setting has been a prolific area of research in educational measurement. Measurement journals such as *Applied Measurement in Education, Applied Psychological Measurement, Educational Measurement: Issues and Practice, and the Journal of Educational Measurement* have published a large number of articles reporting research on standard-setting methods. At national conferences, an even larger number of papers on standard setting have been presented. In 1994, NAGB and NCES organized a joint conference with 19 keynote addresses that offered a comprehensive review of the state of the art in standard setting for large-scale assessments (Bourque, 1995). The vast majority of these papers evaluate the Angoff method, propose a modification of its method, or research a variation in its implementation. The authors of Chapter 5 acknowledge that the Angoff method is the "most prevalent approach to standard setting currently in use" (p. 165). Chapter 5 includes only a few references to journal articles, presentations, and chapters in conference proceedings from this massive body of literature.

Two kinds of references are particularly conspicuous by their absence from the list of references. First, since their initiation of standard setting on NAEP, the staffs of NAGB and its contractors have conducted numerous studies, field-testing many variations and improvements of their standard-setting methods. The nature of these studies is reviewed later in our response. A recent list of related studies compiled by NAGB includes 58 archival publications, 55 technical reports, and 43 conference presentations. Only a small number of the studies are referenced in the NAS Report. In particular, given the extensive criticisms in the chapter on the 1996 standard-setting process, we missed references to "Setting Achievement Levels on the 1996 National Assessment of Educational Progress in Science: Final Report" (ACT. 1997a, 1997b, 1997c, 1997d, 1997e). The five volumes of this report, over 1,200 pages of documentation, describe all procedures, technical decisions, research projects, pilot studies, and results and include all meeting agendas, briefing books, and training material for the setting of achievement levels on the 1996 NAEP Science Assessment. We know that all of these documents were provided by NAGB to the authors of Chapter 5 for their review. Nevertheless, it seems the authors chose to ignore or exclude these and other primary sources in preparing their review.

Second, an important source of information for a professional review is publications in scientific journals. This category is important because the process of independent, blind review through which these journals select their articles embodies the generally recognized principle that critique by peers is an integral part of the process of ensuring scientific quality. Chapter 5 lists no more than 10 titles from this category. Instead, the authors chose to rely almost exclusively on an earlier evaluation of NAEP and the Trial State Assessment by the National Academy of Education (NAE). Nearly half of the references in Chapter 5 are to these National Academy of Education reports or accompanying publications. All of this work was carried out on the 1990 and 1992 standard-setting initiatives. Rather than conducting a scholarly, independent evaluation based on up-to-date primary information and publications that have withstood the critique of fellow researchers, the NAS researchers seem to have chosen to appeal to the authority of NAE, the General Accounting Office (GAO), and a small group of evaluators with well-known convictions about standards and the Angoff method.

Not only are the references in Chapter 5 inadequate, but the chapter is also weak with respect to other aspects of the scientific approach. For example, such an approach would begin with a careful delineation of the object of evaluation, a description of the approach, and an explanation of the evolution over nearly a decade in the methods of standard setting used by NAGB. It would then reach conclusions through a clear chain of logic. Instead, after a handful of introductory pages, the authors summarize a few "selected findings from past NAEP evaluations and research" (p. 166) and then immediately jump to the conclusion that the "achievement levels are flawed" (p. 167).

In addition, the authors have failed to recognize even such a simple thing as the three qualitatively different stages in the NAGB achievement-levels-setting procedure. The procedure starts with a description of the achievement levels for the subject area, which are based on NAGB's policy definitions of the basic, proficient, and advanced levels. In the next stage, panelists are trained and asked to translate the achievement level descriptions into

performance standards on the NAEP score scales. In the last stage, NAGB reviews the previous work and adopts or modifies the recommended achievement levels.

Because Chapter 5 treats all stages as one big stage, it is difficult to associate Chapter 5 conclusions with specific aspects of the standard-setting process. Consider, for example, the 1996 Science achievement levels setting process, which resulted in a very small percent of students in the advanced category. If this result is problematic, is it because of a flaw in the version of the Angoff method used to translate the level descriptions into achievement levels? Or, could it be that the method performed well, but there was a flaw in the procedure used to establish the achievement levels' descriptions? Or, were the science achievement levels' descriptions just too challenging for students? Or, was the NAEP science item pool more difficult than pools for other standard settings?

The same lack of logic is reflected in the casual conclusion that the standard-setting process is flawed because of its sensitivity to item response format (p. 166). Is the procedure responsible for this presumed "flaw"? Or, are there persistent differences in student motivation? Or, differences in item scoring? Could the explanation lie in the fact that the achievement level descriptions are more in line with the scoring rubric of some types of items? Or, could there possibly have been problems in scaling the data? Should our conclusion be that there was no problem at all—the results were a reflection of the way judges reacted to test items? In short, what arguments do the authors have to support their immediate conclusion that the standard-setting method was flawed? There appear to be many possible explanations for the results, and it is not obvious that flaws in standard setting are necessarily the explanation.

We believe that Chapter 5 was based on an approach that is lacking in scientific credibility. As already indicated, Chapter 5 seems at variance with important principles related to the functions of the NAS. Likewise, Chapter 5 does not meet important utility, feasibility, and accuracy standards formulated in the program evaluation standards published by the Joint Committee on Standards for Educational Evaluation (1994). For example, consider Standard A-10, "the conclusions reached in an evaluation should be explicitly justified, so that stakeholders can assess them." This standard and many of the other standards were clearly violated in the NAS evaluation of NAGB's achievement-levels-setting process.

Specific Criticisms

The authors of Chapter 5 concluded that the current standard-setting procedures used with the NAEP assessments are fundamentally flawed. This conclusion was based primarily on three factors. A brief discussion of each of these factors follows.

Believability of the Results

The first reason cited in Chapter 5 in support of the conclusion concerns the believability of the results, particularly the small percentages of students judged to be advanced in science. To support this claim, the authors note that NAGB revised the science advanced achievement level recommended by the achievement-levels-setting panelists, with the result that a larger proportion of students were classified as advanced than would otherwise have been the case. Moreover, the authors claim that numerous external comparison studies conducted by the NAE (e.g., Burstein, Koretz, Linn, Sugrue, Novak, & Baker, 1996; McLaughlin et al., 1993; Shepard, Glaser, Linn, Bohrnstedt, 1993) support the conclusion that NAEP proficient and advanced cutscores are consistently set too high.

However, we have serious concerns about the NAS authors' cursory analysis of data and evidence that leads them to the general conclusion that the NAEP achievement levels place too few students in the advanced levels. The following observations are offered to illustrate our concerns:

1. Based on the occurrence of small proportions of students in some categories (e.g., advanced science), the NAS authors appear to draw the inference that the process of arriving at achievement levels is flawed. We do not concur that the small proportions are evidence that achievement levels are flawed. High achievement levels (which lead to small proportions) are an expected consequence of rigorous achievement-level descriptions. We strongly believe, for example, that the advanced achievement-levels descriptions in science are challenging. High achievement levels and small proportions of student scores above them are both expected and consistent with the descriptions. Consequently, it seems incongruous to criticize the achievement-levels-setting process (i.e., the process of collecting ratings that lead to the achievement levels) as resulting in achievement levels that are too high.

2. The logic proposed by the NAS committee seems to be that "the current procedures are fundamentally flawed" (p. 174), because, in part, "the results are not believable," based on the claim that "too few students are judged to be advanced" relative to other conceptions of *advanced*. This logic is puzzling on three counts. First, it suggests that someone knows, before the fact, something about what the answer (i.e., the standards) should be or what the answer should not be. If the Chapter 5 authors truly believe that the answer can be ascertained, then they should inform NAGB, so NAGB can decide as a matter of policy whether or not to incorporate such preconceptions into the standard-setting process.

The second puzzling aspect of the authors' logic and assertions is the presumption that results from testing programs unrelated to NAEP are an obvious and appropriate benchmark for evaluating NAEP results. Such comparisons are, at best, problematic when the testing programs differ substantially from NAEP programs or have different purposes.

The third puzzling aspect of this logic is the apparent assumption that the word *advanced* means (or should mean) the same thing in different testing programs. In particular, there is no reason to think that the word *Advanced* in the Advanced Placement program has any particular bearing on the word *Advanced* as

a description of one of the achievement levels in NAEP.

3. The authors also fail to acknowledge that an extraordinarily large number of the individuals participating on the various NAEP standard-setting panels have indicated satisfaction with the process and confidence in the results of the process. Since 1992, NAGB has set over 60 achievement levels. The achievement levels range across seven content areas and three grade levels. Within grades and subjects, the achievement levels resulted from the work of standard-setting panels, each of which included approximately 20–30 individuals. Throughout each of the standard-setting sessions, panelists conveyed information about the ALS process through extensive questionnaires. In almost all cases, the panelists indicated strong satisfaction with the process and confidence in the results. There were only three instances in which the results of the standard-setting process were challenged (i.e., the three achievement levels for Grade 8 Science, 1996), and, in those instances, by only a small number of the panelists.

4. The Chapter 5 authors point to the lack of correspondence between NAEP achievement level results and external evidence of student achievement in other contexts (e.g., Advanced Placement examinations/courses) to support their assertion that too few students attain the advanced level on NAEP. Although there are sensible uses of Advanced Placement data in ongoing validation efforts related to the achievement levels, Advanced Placement data do not provide a clear criterion for validating any specific NAEP achievement level. In particular, the content of the Advanced Placement examinations and the testing conditions is very different from the NAEP. Even so, there is some evidence that the percentage of students scoring at high levels on the Advanced Placement science test corresponds reasonably well with the percentage of students at the advanced level on the Grade 12 Science NAEP. In order to provide a context for judging the recommendations for the levels on the 1996 NAEP Science Assessment, NAGB requested information from ETS on the percentage of Grade 12 public school students who would have been eligible for selection into the NAEP sample and who did receive a grade of 3, 4, or 5 on any AP Science test in high school. The College Board database was scanned for the 1994 sophomores, 1995 juniors, or 1996 seniors who met the criteria (without duplication). The final results indicated that only 1.93% of the Grade 12 cohort took an AP science course and received a score of 3, 4, or 5 on the AP high school science test.

5. Linn (1999) reported, for 11 states, the comparisons between the percent of proficient and advanced students identified by NAEP and state assessment findings. For two of the states, students actually appeared to perform better using the NAEP achievement levels than the states' own standards. For four of the other states, the results were relatively close. These findings certainly do not send out shock waves about the invalidity of NAGB's achievement levels because they were set too high. The achievement levels may be high, but Linn's findings do not provide evidence of invalidity. On the contrary, the findings raise more concerns about the performance standards being used in states where the differences from NAEP are very large. For example, in one of the states, NAEP results using the achievement levels indicated that about 15% of the students were performing at the proficient or advanced levels. Using the states' own performance standards, the figure was 90%!

Item Types and Item Difficulty

The second reason given by the authors of Chapter 5 to support the conclusion that NAEP's standard-setting procedures and results are fundamentally flawed is that "the achievement levels-setting results vary significantly depending on the type and difficulty of the items used in the judgment process" (p. 174). Specifically, constructed response items (e.g., short response, extended response, and performance items) typically result in higher standards than those set using multiple-choice items, and the standards obtained using easy items are different from those obtained using difficult items.

In a review of the early NAE studies that, according to the NAS, supports this conclusion, Kane (1993, 1994, 1995) argued, we think persuasively, that such evidence was not necessarily an indictment of the Angoff method. Although the authors of these NAE studies viewed the differences in results based on different item types or item difficulty to be method artifacts that should have been eliminated by the panelists, Kane observed that this was not the only reasonable interpretation. He states:

> One could assume that the apparent difference in the quality of student performance between dichotomous items and extended response items . . . [is] real and that students are meeting judges' expectations, on recognition tasks, but not doing as well, relative to judges expectations, on tasks that require an extended response. . . . The fact is that many scholars believe extended response items tap aspects of student achievement not directly assessed by multiple-choice items. (Kane, 1995, p. 125)

If, in fact, the two types of items are measuring different aspects of student achievement, then different levels of achievement may be expected.

Brennan makes a similar point:

> Some find this [i.e., different performance standards for dichotomous items versus extended response items] to be reasonable, even expected, based on the fact that the two types of items are intended to measure different content and/or constructs. (Brennan, 1998, p. 9)

Brennan further noted that if such differences are to be used as evidence of the invalidity of the standard-setting procedure, then it must be shown "that the scaling procedures used in NAEP should lead to similar results for both types of items and that the lack of congruence is attributable to inadequacies in the standard-setting and not inadequacies in the scaling" (p. 9).

Kane (1995) also raises the scaling issue with respect to the interpretation of differences between the achievement levels set on the basis of either hard or easy items. He notes that to attribute such differences to

the standard-setting method ignores the possibility that:

> the NAEP scaling system, which is probably the most complicated such system in use, [may not be] doing a good enough job in adjusting for item difficulty to ensure that estimates of scores on the IRT-based achievement scale based on the hardest items would generally be equivalent to the estimates based on the easiest items. (p. 127)

Kane continues by noting that:

> Research on this type of vertical scaling has not generally been very encouraging, and, therefore, it would seem prudent to at least consider the possibility that the scaling procedure might be having an impact.... (Kane, 1995, p. 127)

Another factor that must be considered in this discussion is the differential motivation of examinees with respect to the two types of items. It is well documented, for example, that the proportion of students omitting extended response items is greater than the proportion omitting multiple-choice items. Likewise, DeMars (1998) also observed that for low-stakes tests students were more likely to omit extended response items than multiple-choice items.

Difficulty in Estimating Probabilities

The third reason Chapter 5 authors use to support their claim that the procedures are flawed is that panelists have difficulty estimating probabilities for test items (p. 175). These authors agree with a previously made NAE claim that the judgment task posed to raters "is too difficult and confusing" to categorize performance at "three different levels" (p. 166).

There are several problems with this claim, as pointed out in a variety of earlier arguments (e.g. Cizek, 1993; Kane, 1995; Zieky, 1995). First, as recognized by virtually all researchers, including the NAE panel, the Angoff method is the most prevalent approach to standard setting in use (p. 165). The Angoff method is the most prevalent approach because it generally has been judged to be the best approach. For example, Jaeger (1990) stated that:

> There appears to be a developing consensus, ... that Angoff's procedure produces more reasonable standards than do its competitors. Based on very limited results, it appears that Angoff's procedure will often produce more stable (and hence more reliable) standards than will its competitors. (p. 19)

Others have taken similar positions favoring the Angoff approach. Mehrens (1995), in a detailed review of various standard-setting approaches, concluded that the Angoff method is regarded as the preferred model. It typically provides a reasonable standard with good psychometric properties.

To conclude, based on two references (see p. 175), that panelists cannot do something that panelists have been doing for decades and that research supports they have been doing quite well seems inappropriate. Further, such a conclusion is contrary to evidence collected from NAEP standard-setting panelists who consistently indicate that they understood the task. For example, panelists in the mathematics, reading, and writing achievement-levels-setting process stated that they were confident, they understood the task, they were not coerced into a decision, and they felt the results were credible (ACT, 1993b, September).

This is not to suggest that standard setting is easy. Making judgments about levels of performance that correspond to advanced, proficient, and basic will be difficult no matter what specific procedure is implemented. Certainly there is evidence in the literature that making probability estimates is difficult. However, the particular cognitive decision being made in the standard-setting process is one that panelists have been thoroughly trained to do. The standard-setting process is a 5-day process that involves thorough training and iteration of tasks with feedback to the panelists between the iterations. The design and implementation of the process is largely aimed at helping the panelists with their tasks, and there is considerable evidence that panelists can perform their tasks very well. For example, as discussed later in this article, in each grade of every standard-setting process, two groups of panelists work independently. Comparisons of these mini-replications do not suggest that panelists are behaving irrationally or erratically. On the contrary, these comparisons strongly suggest that the panelists are quite consistent in their final judgments.

In discussing the cognitive complexity of the task, it should be mentioned that there is legitimate debate about just how panelists interpret their task. Zieky, for example, suggests the following interpretation:

> I don't have any research to verify this, but my experience in standard setting leads me to believe that judges involved in the major item-judgment methods of setting standards are not really making estimates of the performance of some hypothetical group of minimally competent examinees.
>
> I think the actual process is closer to one in which a judge looks at an item and decides that he or she would not be willing to call an examinee minimally competent unless the examinee had a chance of at least x of getting the item right. I believe that judges are not making sloppy estimates of probability. They are directly expressing their own values about what level of performance they would consider good enough to call minimally competent. (Zieky, 1995, p. 30)

The accuracy of the NAE claim that the Angoff method requires an impossible cognitive task was directly addressed by Cizek. Based on his analyses, Cizek concluded that:

> Put simply, there is no evidence in the psychometric literature to support the Report's contention that the Angoff method requires an "impossible cognitive task." Indeed, the literature presents precisely the opposite conclusion ... (Cizek, 1993, p. 10)

Kane also offered rebuttals to several studies related to the Chapter 5 authors' conclusion regarding the difficulty of the judgments and concluded as follows:

> The evidence developed in the five studies of the technical properties of the 1992 NAEP standard setting does not seem to justify the conclusion (Shepard et al.,

> 1993, p. 77) based largely on these studies, "that the Angoff procedure is fundamentally flawed because it depends on cognitive judgments that are virtually impossible to make." (Kane, 1995, p. 129)

The authors of Chapter 5 also suggest that panelists cannot judge the relative difficulty of items. However, correlations between totally independent Angoff ratings and item difficulties are substantial (see, e.g., Hambleton & Bourque, 1991). Further, item judgments are averaged for each panelist, and the final performance standards are averaged over panelists. The achievement levels are not dependent on the judgment of a single item by a single judge. Rather, for example, setting standards on a 50-item test with a panel of 15 judges would produce 750 item ratings to be averaged for each round of ratings. The NAS authors' concern about the ratings of single items is both unwarranted and misleading. It would be comparable to making an important decision about examinees based on their performance on a single item.

In determining whether a task can be done, one reasonable source of information is the people who have been asked to do the task (see McLaughlin, 1993a). Kane addressed this issue as follows:

> Nevertheless, the judges asked to complete Angoff ratings managed to do so without complaint in both NAGB standard-setting efforts and the NAE replication (McLaughlin, 1993b) of one of the NAGB studies. The results for NAE's replication were very close to the original NAGB results, and both of these sets of results were in general agreement with the results of NAE's whole book ratings (McLaughlin, et al., 1993c). Furthermore, the Angoff method has been used to set passing scores on a host of licensure and certification tests, as well as on numerous state testing programs, without major complaints from the judges involved. Not only can judges do the task, but they also do not seem to find it particularly anxiety provoking as such (although they do sometimes find it tedious). (Kane, 1995, p. 124)

Thus, the conclusion of the authors that "the judgement tasks are difficult and confusing" (p. 182) does not seem to be true.

At various places in Chapter 5 (most notably in the last full paragraph on p. 175) statements are made that suggest the authors believe that those involved in NAEP standard-setting have not considered alternative methods. This is not true. Many alternative procedures have been considered, and several have been pilot tested. For example, Kane (1993, 1995) provided some compelling arguments against the contrasting groups method that has been proposed by some of the critics of current approaches. Also, an item mapping procedure has been considered, but it requires the arbitrary selection of a response probability (e.g., 50%, 65%, 74%) that may substantially affect results (ACT, 1997b). In addition, holistic procedures have been pilot tested, but those procedures generally have yielded higher, not lower, achievement levels (ACT, 1997d). Additional details about some of NAEP's exploration of alternative procedures are provided in the next section.

Integrity of the Achievement-Levels-Setting Process

Chapter 5 leaves the oversimplified impression that the achievement-levels-setting (ALS) process is a single event. In fact, the ALS process begins with the development of policy definitions for the achievement levels and ends when NAGB accepts the cutscores on the NAEP scale and exemplar items for use in reporting NAEP results. In between these two policy components, achievement level content descriptions are developed, a formal standard setting methodology is implemented, and statistical procedures are used to compute cutscores on the NAEP reporting scale. Further, the entire process is thoroughly reviewed and evaluated.

The ALS is not only a multiphase process but also a dynamic one. With each implementation, components of the process are refined. For example, prior to the science ALS, the panel that applied the standard-setting methodology also developed the achievement level descriptions (ALDs). For the current process, ALDs are developed along with the NAEP framework. Both the frameworks and the ALDs receive NAGB policy approval before they are used to guide the formal ALS methodology.

Achievement levels setting is a judgmental process, but that does not mean that the process is arbitrary. The judgment task should be well defined, and it should be performed by qualified people who are provided with the information and training they need to do the task. The process used for setting achievement levels for the NAEP has been designed to meet these criteria.

The Panelists

The panelists involved in the process have been carefully selected because of their knowledge of the content, their familiarity with the student population, and their reputation within the educational community. The procedure used for selecting panelists is replicable. Nominators are identified through a stratified random sampling process, ensuring both that the panelists represent all regions and specified demographic attributes within the country and that multiple samples of panelists can be obtained for use in research and pilot studies. The methods used to identify and select qualified panelists are well documented in ACT reports (e.g., ACT, 1997b).

The panelists consist of teachers at the specific grade level, nonteacher educators, and members of the general public with relevant background and/or experience in the subject matter tested. The use of a broadly representative group of panelists is not only a matter of NAGB policy but also a matter of legislative requirement. Recommendations made by the authors of Chapter 5 (see Appendix D in the NAS report) seem to represent at least a partial challenge to both the law and NAGB policy. However, they do not provide evidence or a compelling argument supporting an alternative policy. Further, ACT reports of ALS processes throughout this decade indicate that the three types of panelists tend to set quite similar standards.

The Task

The judgmental task panelists perform was designed through a contin-

uing program of research and refinement. The entire process is conducted under the thorough review of a Technical Advisory Committee for Standard Setting (TACSS), composed of highly experienced persons in the fields of psychometric theory, educational policy, and standard setting. The current effort to set achievement levels for NAEP writing demonstrates the careful approach that is used for the ALS process. The TACSS did not believe that the achievement levels developed for 1992 NAEP Writing met appropriate standards of quality. They recommended to NAGB that the results of that ALS process not be used for reporting (correspondence between ACT and NAGB, June 14, 1993). NAGB accepted that recommendation. A refined ALS process was developed and applied to the revised 1998 NAEP Writing assessment. The results of that process have been accepted by NAGB and are now being used to report the 1998 NAEP Writing results. This example shows that the results of the ALS process are not given automatic approval but carefully reviewed and used only when they meet high professional standards.

The evolution of the ALS process began with the work of Hambleton and Bourque (1991). Refinement of the process has continued since 1992 when ACT became the NAGB ALS contractor (ACT, 1993b). Numerous field trials, pilot studies, and experience with operational implementation for NAEP tests in mathematics, reading, writing, geography, U.S. History, and science have guided the refinement process. The result is the current process for NAEP civics and writing. The details of the process can be found in a series of reports (ACT, 1997f, 1997g, 1997h, 1998).

The evolutionary character of the ALS process has not been acknowledged in the NAE report. Chapter 5 references the 1993 NAE evaluation (Glaser, Linn, & Bohrnstedt, 1993) as a description of the ALS process. Since the NAE evaluation, the 1994 NAEP Geography and U.S. History processes have been implemented, leading to the process used for 1996 NAEP Science, the focus of Chapter 5. The ALS process used for NAEP Science contained refinements that addressed many of the points raised in the NAE report. For example, whole booklet feedback was implemented to provide judges with holistic information about student performance. Other changes were made for 1996 NAEP Science because of the large number of performance items. The reference sources given in Chapter 5 do not acknowledge these differences.

The ALS process has been very public, and it has been subjected to many formal evaluations. In fact, it is probably the most carefully considered standard-setting process in the country, if not the world. The openness of the process and the critical evaluations that have taken place have served as stimuli for extended research on standard-setting processes. As documented in ACT (1995), replicability of results, psychometric soundness, and reasonableness of standards have been investigated for numerous standard-setting processes. For example, whole booklet classification methods, student performance classifications, item mapping, and item response string estimation techniques have been studied as alternatives to the current methodologies. In all cases, the alternative methods lacked psychometric soundness and/or yielded results that appeared highly questionable. It is not that ACT and NAGB have ignored the evaluations of the ALS process; they have used the evaluations to guide the refinement of the basic process and the search for alternatives. Overall, evaluations of alternatives have shown them to have less desirable characteristics than the current ALS process.

Variations on the current process have also been evaluated. An array of computer simulations, field tests, and pilot test studies have been used for these evaluations. Possible modifications to standard-setting procedures, feedback mechanisms, and training approaches have been considered. For example, two pilot studies were conducted prior to the operational ALS process for 1996 NAEP Science. These studies focused on the procedures needed for the hands-on science tasks, several response probability criteria for item mapping, and the effect of consequences information on the ratings. Reckase and Bay (1998) provide a summary of the full scope of the ALS research efforts.

A unique feature of the ALS process is that it includes replications of the rating component and the computation of the cutscores. The achievement level cutscores are determined for two groups working independently at each grade level. The similarity of the achievement level cutscores estimated for the two replications has been uniformly high for all of the ALS projects. The pilot testing processes have also yielded partial replications with different groups and at different times. These replications show that the achievement level cutscores consistently have small standard errors.

Training

Field tests and pilot studies are also used to verify the effectiveness of training procedures and the 5-day ALS process. For 1992 NAEP mathematics, panelists received about 7 hours of training concerning NAEP, NAGB policy, the achievement levels, and the ALS process (ACT, 1993a). The current process begins training with extensive briefing materials sent in advance of the ALS session and continues with 2 days of training about NAEP, NAEP scoring, NAGB policy, the ALS process, and the feedback the panelists receive (ACT, 1998). The form of training has also been refined, now using multimedia presentations and carefully planned instructional tasks.

Summary

The end result of this development work is an ALS process that has been carefully designed and refined, is carefully reviewed before operational implementation, and is thoroughly evaluated. Chapter 5 does not reflect the careful planning, extensive research support, and thorough review of this process.

Conclusion

Chapter 5 and Appendix D in the NAS Report present a disconcerting juxtaposition of perspectives. Chapter 5 aims at characterizing NAGB's achievement-levels-setting process as "fundamentally flawed" with hardly a reference to, not to

mention a discussion of, numerous reports and papers that document the process and challenge this rhetoric. Basically, Chapter 5 devotes two lines to mentioning alternative perspectives. Then, Appendix D devotes almost six pages to suggesting a very complicated process that has never been subjected to even a single pilot study. It is trivially easy to criticize an extant process if one ignores documentation and evidence, and it is equally easy to assign potential and virtue to a new process if it is not subjected to a reality test. Both perspectives, and particularly their juxtaposition, seem to us to be misinformed, misleading, and basically unsound from a scientific perspective. In particular, the "fundamentally flawed" rhetoric that pervades *Grading the Nation's Report Card* is not just unwarranted—it is simply wrong. No standard-setting process is, or could be, perfect, but the ACT/NAGB process has much to recommend it.

In this article, we have provided a detailed response to many aspects of the review/evaluation of NAGB's initiatives to set achievement levels on the NAEP. Our review of the evidence suggests that Chapter 5 on "Setting Reasonable and Useful Performance Standards" for NAEP constitutes a one-sided, incomplete and inaccurate accounting of the NAGB/ACT standard settings conducted during this decade. As such, Chapter 5 of the NAS Report is a disservice to NAGB, educational policymakers, educators, and the public. In our opinion, in this instance, the NAS review and screening process clearly failed to produce a trustworthy and credible evaluation.

Notes

The authors serve as technical advisors to ACT on the NAEP standard-setting process. They are grateful to Susan Loomis, of ACT, and Mary Lyn Bourque, of the National Assessment Governing Board, for providing the technical assistance needed in the preparation of this article.

References

ACT. (1993a). *Description of mathematics achievement levels-setting process and proposed achievement level descriptions: Volume 1.* Iowa City, IA: Author.

ACT. (1993b). *Setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading, and writing: a technical report on reliability and validity.* Iowa City, IA: Author.

ACT. (1995, December). *Research studies on the achievement levels set for the 1994 NAEP in geography and U.S. history.* Iowa City, IA: Author.

ACT. (1997a). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science: final report volume I, pilot study 1.* Iowa City, IA: Author.

ACT. (1997b). *Setting achievement levels on the 1996 National Assessment of Educational Progress in Science: final report volume II, pilot study 2.* Iowa City, IA: Author.

ACT. (1997c). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science: final report volume III—achievement levels setting study.* Iowa City, IA: Author.

ACT. (1997d). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science: final report volume IV, validity evidence and special studies.* Iowa City, IA: Author.

ACT. (1997e). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science: final report volume V, technical decisions and NAGB actions.* Iowa City, IA: Author.

ACT. (1997f, March). *A proposal to the National Assessment Governing Board for developing achievement levels on the 1998 NAEP in civics and writing: Technical Proposal.* Iowa City, IA: Author.

ACT. (1997g, September). *Developing achievement levels on the 1998 NAEP in civics and writing: planning document.* Iowa City, IA: Author.

ACT. (1997h, December). *Developing achievement levels on the 1998 NAEP in civics and writing: design document.* Iowa City, IA: Author.

ACT. (1998, November). *Developing achievement levels on the 1998 NAEP in civics and writing: planning document, update 1.* Iowa City, IA: Author.

Bourque, M. L. (Ed.). (1995). *Proceedings of the joint conference on standard setting for large-scale assessments.* Washington, DC: National Assessment Governing Board, National Center for Education Statistics.

Bracey, G. W. (1999). The ninth Bracey report on the condition of public education. *Phi Delta Kappan, 81* (2), 147–168.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17*(1), 5–9, 30.

Burstein, L., Koretz, D. Linn, R., Sugrue, B., Novak, J., & Baker, E. L. (1996). Describing performance standards: Validity of the 1992 National Assessment of Educational Progress achievement level descriptions as characterizations of mathematics performance. *Educational Assessment,* 3(1), 9–51.

Cizek, G. J. (1993). *Reaction to National Academy of Education Report, "Setting Performance Standards for Student Achievement."* Washington, DC: National Assessment Governing Board.

DeMars, C. (1998). *The impact of test consequences and response format on performance.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement.* Stanford: National Academy of Education.

Hambleton, R. K., & Bourque, M. L. (1991). *The levels of mathematics achievement* (Vol. III, Tech. Rep.). Washington, DC: National Assessment Governing Board.

Jaeger, R. M. (1990). Establishing standards for teacher certification tests. *Educational Measurement: Issues and Practice, 9* (4), 15–20.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs* (2nd ed). Thousand Oaks, CA: Sage.

Kane, M. (1993). *Comment on the NAE evaluation of the NAGB achievement levels.* Washington, DC: National Assessment Governing Board.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425–461.

Kane, M. (1995). Examinee-centered vs. task-centered standard setting. In M. L. Bourque (Ed.), *Joint Conference on Standard Setting for Large-Scale Assessments.* Washington: NCES-NAGB.

Linn, R. L. (1999, February). *Assessment and accountability systems.* Paper (invited) presented at the meeting of the American Association of School Administrators, New Orleans.

McLaughlin, D. H. (1993a). Validity of the 1992 NAEP achievement-level setting process. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Setting performance standards for student achievement: Background studies.* Stanford: National Academy of Education.

McLaughlin, D. H. (1993b). Order of Angoff ratings in multiple simultaneous

standards. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Setting performance standards for student achievement: Background studies*. Stanford: National Academy of Education.
McLaughlin, D. H. (1993c). Rated achievement levels of completed NAEP mathematics booklets. In R. Glaser, R. L. Linn, & G. Bohrnstedt (Eds.), *Setting performance standards for student achievement: Background studies*. Stanford: National Academy of Education.
McLaughlin, D. H., DuBois, P. A., Eaton, M. S., Ehrlich, D. E., Stancavage, F. B., O'Donnell, C. A., Yu, J. Y., DeStefano, L., Pearson, D., Bottomley, D., Bullock, C. A., Hanson, M., & Rucinski, C. (1993). Comparison of teachers' and researchers' ratings for students' performance in mathematics and reading with NAEP measurement of achievement levels. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Setting performance standards for student achievement: Background studies*. Stanford: National Academy of Education.
Mehrens, W. A. (1995). Methodological issues in standard setting for educational exams. In M. L. Bourque (Ed.), *Proceedings of the joint conference on standard setting for large-scale assessments* (pp. 221–263). Washington, DC: National Assessment Governing Board, National Center for Education Statistics.
National Academy of Sciences. (1998, July). *A unique national resource*. Washington, DC: Author.
Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). *Grading the nation's report card*. Washington, DC: National Academy Press.
Reckase, M. D., & Bay, L. (1998, April). *Analysis of methods for collecting test-based judgements*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego.
Rothstein, R. (1998). *The way we were? The myths and realities of America's student achievement*. New York: Century Foundation.
Shepard, L. A., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement* (Final Report). Stanford: National Academy of Education.
Zieky, M. (1995). A historical perspective on setting standards. In M. L. Bourque (Ed.), *Joint Conference on Standard Setting for Large-Scale Assessments* (pp. 1–38). Washington, DC: NCES-NAGB.

A Response to ACT's Technical Advisers on NAEP Standard Setting

James W. Pellegrino
Vanderbilt University

My colleagues and I who were members of the National Research Council (NRC) Committee on the Evaluation of National and State Assessments of Educational Progress appreciate the interest and extraordinary effort of Hambleton et al. in analyzing chapter 5 of *Grading the Nation's Report Card*. The authors, as respected researchers and as members of the Technical Advisory Committee on Standard Setting (TACSS) to American College Testing (ACT), Inc., the National Assessment Governing Board's (NAGB) contractor for NAEP standard setting, have made significant contributions to research and development in educational standard setting. They have a legitimate interest and stake in the conclusions and recommendations of the NRC Committee's report.

With the publication of *Grading the Nation's Report Card*, the Committee on the Evaluation of National and State Assessments of Educational Progress completed its task, so it is not in a position to comment in detail on the TACSS response. The NRC Committee's report and TACSS' response stand as contributions to the ongoing policy debates and research work on NAEP.

James W. Pellegrino is Frank W. Mayborn Professor of Cognitive Studies, Department of Psychology and Human Development & Learning Technology Center, Peabody College of Vanderbilt University, Box 45, Peabody Station, Nashville, TN 37203. His specializations are human cognition, cognitive development, instructional technology, and applications of cognitive theory to issues in assessment and instructional design.