



National Assessment Governing Board

National Assessment of Educational Progress

STATEMENT BY MARK D. MUSICK, CHAIRMAN NATIONAL ASSESSMENT GOVERNING BOARD

Comments on National Academy of Education Panel Report on NAEP Achievement Levels

Over the past four years the National Assessment Governing Board has made a major effort to carry out one of the key responsibilities given it by law: to identify "appropriate achievement goals for each...grade (and) subject area to be tested under the National Assessment."

We have consulted widely and listened carefully--both to the public and to expert opinion. To conduct the process for setting achievement levels for the 1992 National Assessment, we contracted with American College Testing (ACT), which has extensive experience setting standards in many fields.

We selected a method for standard-setting that is the nation's most widely-used--from certifying surgeons, auto mechanics, and food inspectors to setting graduation and promotion standards in colleges and schools. We took into consideration the best technical advice we could find. This included a panel of leading experts in measurement and standard-setting, convened by ACT. These experts believe strongly in the soundness of the process used.

As the National Academy of Education Panel notes in its report, setting standards is fundamentally a matter of policy judgment, not a scientific search for "truth." That judgment should be informed and deliberate--and the process the Governing Board used certainly was. The process is described fully in a report by ACT, which we urge you to read. We are convinced that with its carefulness and complete disclosure, the process fully meets the Standards for Educational and Psychological Testing, established by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

Are the standards reasonable or are they "too high?" We acknowledge setting high standards for the National Assessment; we believe high standards are important for the long-term future of our nation. Indeed, the central idea in defining the proficient level for each grade is that students demonstrate "competency over challenging subject matter." This language is borrowed from the National Education Goals. We deliberately did not seek to ratify the status quo.

800 North Capitol Street, N.W.
Suite 825
Mailstop 7583
Washington, D.C. 20002-4233
(202) 357-6938

The shortcomings of that status quo were illustrated well by the National Adult Literacy Survey, which recently reported about 90 million Americans functioning at low levels in reading and mathematics. The performance of young adults, it said, had declined since 1985. As Secretary Riley noted, the report was a "wake-up call to the sheer magnitude" of the literacy problems in our country.

These results make it even more difficult to believe that the NAEP achievement levels are "too high." How can we suggest lowering standards for young people when nearly one-half of America's adults do not read well enough or apply math well enough?

Might the process of setting the achievement levels be improved? Yes, of course it can be.

Will the process change as the National Assessment of Educational Progress evolves? Certainly. The Board is open to proposals for new standard setting techniques and will consider them in setting achievement levels in other subjects. Many of the suggestions made by the Academy panel should be considered.

Yet, unlike the panel, we believe standards were set reasonably on NAEP this year, using a method employed on many tests for more than two decades. The panel contends that this system of assembling a range of opinions to set a "passing" score--not the execution of it--is "fundamentally flawed." Yet, the method, named for its developer William Angoff, has withstood legal challenge and has served well on examinations with much greater consequences for individuals than the National Assessment.

Not only is the Angoff method supported by extensive research, but a number of the Academy's own research studies confirm the validity and reliability of the NAEP achievement levels. Yet, these results are passed over in the panel's report.

In place of the standard-setting method used for NAEP the panel offers no clear alternative and no clear time-table for providing one. Instead, it speaks of first establishing a new "knowledge base" that will take "some years" to develop. Despite its affirmation that standards should be set, its recommendations are so complex and ill-defined that actually developing them would take many, many years, and with no guarantee of success.

In the long interim, a period that apparently extends beyond the year 2000, the report suggests that the Board pick percentiles --perhaps 95th, 75th, and 25th--as "benchmarks" toward which more students should strive. Of course, these would not really be standards at all, just arbitrarily-selected points on the curve of student performance.

Unfortunately, in many respects the panel's report reads like a lawyer's brief rather than a balanced account based on research the panel commissioned. If we follow its recommendations, this country would have no way of measuring progress on academic achievement for the National Education Goals--certainly no way before the year 2000.

Research in setting standards is desirable and must be pursued. But the search for improvement--and criticism that perfection has not been reached--should not be allowed to overturn sound standards arrived at through reasonable, widely-used means.

For this to happen would be a major step backwards for the National Assessment of Educational Progress. It would also set back the efforts by many states and school districts to transform their own testing programs from norm-referenced to standards-based exams--to move away from "national norms" that may well be inadequate. It probably would also delay development of strong national performance standards, a task in which the National Assessment should certainly play an important role.

#

**Reactions to National Academy of Education Report,
"Setting Performance Standards for Student Achievement"**

**Gregory J. Cizek, PhD
Assistant Professor of Educational Research and Measurement
University of Toledo
Toledo, Ohio**

August 1993

**Reactions to National Academy of Education Report,
"Setting Performance Standards for Student Achievement"**

**Gregory J. Cizek, PhD
Assistant Professor of Educational Research and Measurement
University of Toledo
Toledo, Ohio**

August 1993

BACKGROUND

The report of the National Academy of Education (NAE) entitled "Setting Performance Standards for Student Achievement" (hereafter "Report") provides an interpretation of NAE investigations into the procedures surrounding establishment of achievement levels for the National Assessment of Educational Progress (NAEP). The Report synthesizes the investigations of other research efforts commissioned by the NAE Panel on the Evaluation of the NAEP Trial State Assessment. The Report provides a few positive comments and a few constructive recommendations proposing research efforts that may serve to strengthen the NAEP and the achievement levels in the long term. The Report also presents many negative judgments about the process utilized in the establishment of achievement levels.

PURPOSE

The purpose of the present review is to evaluate the conceptual and substantive adequacy of the NAE Report. In preparing this review, the following documents were used:

- 1) "Setting Performance Standards for Student Achievement" (National Academy of Education, 1993) (including 10 studies commissioned by the NAE in preparing the Report);

- 2) "Setting Achievement Levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing" (American College Testing, 1991);
- 3) "NAGB Policy Framework and Technical Procedures for Setting Appropriate Achievement Levels for the National Assessment of Educational Progress" (National Assessment Governing Board, 1991); and,
- 4) "The Reliability and Validity of the 1992 NAEP Achievement Levels" (American College Testing, 1993);

It is the conclusion of my review that the NAE evaluation presents a seriously inaccurate representation of the technical and procedural propriety of the NAEP levels-setting process. In my opinion, the inaccuracies and errors in both factual information and conclusions contained in the NAE evaluation are so substantial as to seriously weaken the document as a reference to inform discussions of the NAEP level-setting process.

The following sections of this review address four major purposes. First, the errors and inaccuracies in the NAE Report are categorized, summarized, and contrasted with the available evidence. Second, the accurate conclusions of the Report are summarized and explicated. Third, a section analyzing the recommendations of the NAE Report is provided. Finally, a summary section including some general observations is included.

I. ERRORS AND INACCURACIES IN THE NAE REPORT

The NAE Report contains an unacceptable quantity of serious errors and inaccuracies in conceptualization, substance, and interpretation. The errors and inaccuracies appear to be the result of several related factors. These errors are described under the following four general headings, which organize this section of my review:

- A) Authority and credibility of the evaluation
- B) Conclusions regarding the use of the Angoff method
- C) Conclusions regarding procedural propriety and results
- D) Conclusions regarding the validity of the NAEP levels.

A. Authority and Credibility of the Evaluation

A serious deficiency of the NAE Report is its foundation in accepted psychometric literature and practice. Because the NAE Report was intended to provide an evaluation of the process used by the National Assessment Governing Board (NAGB) in establishing achievement levels for the National Assessment of Educational Progress (NAEP), it seems imperative that the evaluation be: 1) grounded in relevant psychometric guidelines; and 2) based upon the expertise of recognized experts in the area of standard setting. The NAE Report falls short on both of these criteria.

Authority of guidelines used in the NAE evaluation

The NAE Report does not apply even the most rudimentary guidelines for evaluating test development, administration, and reporting procedures promulgated by the professional testing community.

At minimum, a credible evaluation of any standard setting process would be thoroughly grounded in the universally accepted Standards for Educational and Psychological Testing (AERA/APA/NCME, 1985) or, to a lesser degree, in other guiding documents such as the Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 1988), the primer on standard setting Passing Scores (Livingston & Zieky, 1982), or the chapter on the "Certification of Student Competence" (Jaeger, 1988) found in the professional reference text Educational Measurement. Of these sources, the Standards for Educational and Psychological Testing represents the only compilation, by the major professional groups in testing and measurement, of principles by which test development procedures such as standard setting can be evaluated.

Although a recognized weakness of the Standards is that they do not contain a specific section dealing with setting cutting scores (see Jaeger, 1991), there is guidance on that topic contained within the Standards, with at least six specific guidelines presented in its chapters. There also exists an abundance of professional literature on the subject of standard setting, including the selection and training of judges, implementation procedures for various methodologies, methods of adjusting passing scores, and methods for reporting results.

My review finds the NAE Report to be written almost without recognition of the large body of relevant standard setting literature that currently exists. Lacking this foundation, it is a natural result that the Report did not adhere to the frameworks for evaluating standard setting procedures that are common in the psychometric community. Further, the Report is not structured so as to correspond to any of the professional guidelines codified in the Standards, nor is there any substantial recognition of the authority of the Standards in guiding the NAE evaluation, nor is there even so much as a single reference to the Standards. To put this error in perspective, one might imagine evaluating Christian theology without reference to the Bible or evaluating spelling and usage without reference to a dictionary. The error is especially troubling because the NAE project chairmen and principal investigator are listed in the Standards as key contributors to their development.

That the NAE Report generally avoids grounding in the relevant psychometric literature, and ignores the relevant professional standards is a serious and consequential error. Such an omission cannot simply be dismissed as gross ignorance or careless oversight, but reveals an unwillingness to apply accepted psychometric guidelines to the evaluation of the NAEP levels-setting process. The result of this error is that the evaluators used subjective, unrecognized, and professionally unaccepted standards of evidence in the evaluation. By applying novel standards to guide the evaluation, the NAE Report has not only departed from accepted evaluation practice, but has also failed to: 1) explicitly state the standards that are used; 2) provide a defensible rationale for why accepted standards were not used; and 3) subject the novel evaluation approach employed to the scrutiny of evaluation professionals.

Credibility of the evaluation

In addition to departing from professionally accepted evaluation criteria, the NAE Report and its conclusions are weakened by the fact that the evaluation studies were not conducted by recognized experts in standard setting. Certainly the NAE Report's principal author possesses considerable expertise in psychometrics and testing policy. However, my review of the 10 studies upon which the NAE Report is ostensibly based reveals that none of the principal investigators or graduate assistants who authored the 10 studies even assert that standard setting methodology is among their areas of expertise. It is apparent that many of these individuals possess relevant experience and content area expertise in their disciplines, e.g., reading, mathematics, and so forth. However, such expertise is neither germane nor particularly well suited to providing analysis of the NAEP levels-setting process.

The NAE Report does not describe whether experts in standard setting were solicited to participate in preparing or reviewing the Report, or whether standard setting specialists were asked to conduct any of the 10 commissioned studies. It is possible, given the time interval between when studies could be commissioned and when final study reports were needed, that suitably qualified standard setting experts could not be commissioned. This is unfortunate.

B) Conclusions Regarding the Use of the Angoff Method

The conclusions presented in the NAE Report regarding the use of the Angoff (1971) procedure for establishing achievement levels for the NAEP are largely insupportable and unacceptable from a psychometric perspective. In particular, the conclusion that the Angoff procedure is "fundamentally flawed" as a standard setting procedure is wholly unsupported by the professional literature. Indeed, the literature of standard setting provides ample documentation that the Angoff method is a reasonable, useful, acceptable, and--in many circumstances--preferable method for deriving cutting scores. Further, in a later section of this review, I demonstrate that alternative procedures proposed in the Report are psychometrically indefensible, inconsistent with the objections raised in the NAE Report, and potentially harmful.

Appropriateness of the Angoff methodology

The various standard setting methodologies that exist have been roughly classified as "relative" and "absolute" methods. (An additional category of methods called "compromise" methods exists, though I will not describe them here.)

The relative methods are perhaps the most familiar to the general public. These methods establish a cut-off in a norm-referenced fashion. For example, it might be decided that the top 25% of students should "pass" or receive an "A." Using this type of methodology, the cutting score is set so that the top 25% "pass" regardless of their actual mastery, knowledge, or skill; in simple terms, the top 25% may not know very much at all, but they have at least performed better than the other 75%. This method of setting passing scores has become quite rare in educational settings, possibly because it is so difficult to defend.

The Angoff method, which is undoubtedly the most commonly used method, is similar to many of the other "absolute" passing score methodologies, in that it involves explicit consideration of the knowledge, skills, and abilities of the test takers, the desired levels of knowledge and skill, as well a review of the actual test items that will be used to assess the knowledge and skills of the examinees. Panels of standard setting participants (called "judges") generate estimates of the proportion of examinees who should answer test items correctly. Angoff suggested that:

"a systematic procedure for deciding on the minimum raw scores for passing and honors might be developed as follows: keeping the hypothetical 'minimally acceptable person' in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score earned by the 'minimally acceptable person'" (Angoff, 1971, pp. 514-515).

In practice, a footnoted variation to the procedure Angoff originally proposed has dominated applications of the Angoff method:

"A slight variation of this procedure is to ask each judge to state the probability that the 'minimally acceptable person' would answer each item correctly. In effect, judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities would then represent the minimally acceptable score" (Angoff, 1971, p. 515).

The central conclusion of the NAE Report is that "the Angoff procedure is fundamentally flawed for the setting of achievement levels" (p. xii). This assertion is both far out of line with the preponderance of research evidence and expert opinion in the psychometric community, and unsupported by the evidence presented in the NAE Report itself.

On what basis then does the NAE Report find the method to be "fundamentally flawed?" The following paragraphs recount some of the Report's objections to the use of the Angoff methodology and evaluate the evidence for those objections.

Objection 1 - The Angoff method requires an impossible cognitive task.

The NAE report claims that "the Angoff method requires an impossible cognitive task [of participants in the procedure]" (p. 55). This claim is the central reason for the NAE objection to the use achievement levels for the NAEP, and is the foundation upon which most of the Report's other claims rest. Accordingly, if it can be demonstrated that the central claim is false, related objections would be rendered irrelevant.

The following evidence documents that the NAE Report's central claim is contrary to a wealth of research and experience with the Angoff method accumulated over the past two decades. Further, the claim is inconsistent even with the data and interpretations of NAE's own commissioned studies.

EVIDENCE: The Angoff method is widely accepted and praised within the community of standard setting specialists as a valid method for establishing cutting scores. The weight of the evidence accumulated from research by measurement experts comparing the results of the various item-judgment based procedures is compelling: The Angoff approach seems to be the preferred absolute standard-setting methodology by several criteria.

The following quotations from a small sample of the weighty body of research based evidence illustrate that the NAE Report's conclusion about the Angoff method is strongly contradicted by experts in the field of standard setting:

Mills and Melican (1988) report that "the Angoff method appears to be the most widely used. The method is not difficult to explain and data collection and analysis are simpler than for other methods in this category" (p. 272). Klein (1984) noted that the Angoff method is preferable "because it can be explained and implemented relatively easily" (p. 2). Rock, Davis and Werts (1980) concluded that "the Angoff cutting score seems to be somewhat closer to the 'mark'" (p. 15). Colton and Hecht (1981), in their comparison of the Angoff, Ebel, and Nedelsky methodologies, report that "the Angoff technique and the Angoff consensus techniques are superior to the others" (p. 15). Cross, et al., (1984) concluded that the Angoff method "yielded the most defensible standards" (p. 113). Berk (1986) advised that "the Angoff method appears to offer the best balance between technical adequacy and practicability" (p. 147). Finally, Meskauskas (1986) concluded that, "the present method of choice for standard-setting is the Angoff method" (p. 199).

Although the references listed above describe the general view of the Angoff procedure within the psychometric community, the specific Angoff methodology utilized for the NAEP achievement-levels setting task was also subject to scrutiny by psychometric experts prior to implementation. Information and technical documentation related to the extensive psychometric

review is referenced in the NAE Report (see footnote 25 in the Report) but it is only briefly noted that "details of the procedures...were specified in advance and reviewed by advisory committees" (p. 30). The extensive psychometric review recorded in the technical documentation provides further evidence related to the professional approval of the Angoff procedure as implemented in the NAEP levels-setting process.

Thus, the Report's opinion about the Angoff methodology generally, or about the Angoff methodology as implemented in the NAEP levels-setting process is neither founded in nor supported by the relevant psychometric literature and documentation. Additional examination of the Report's central claim reveals a serious error in reporting the findings of the studies commissioned to support the opinion. Specifically, the central claim that the Angoff methodology presents an "impossible cognitive task" rests on speculation that the provision of three Angoff ratings (i.e., Basic, Proficient, Advanced) for a single item is an overly taxing departure from the usual Angoff procedure in which only one rating is generated. The single study that bears directly on this issue was commissioned by the NAE and is reported in McLaughlin (1993a). The clearly stated finding of this study--not highlighted in the NAE report--is presented below:

"The major finding from this study is that the use of the three-ratings-per-item did not introduce substantial and statistically significant artifactual regularity into the cutpoints...Therefore, it does not seem warranted, based on these results, to recommend against using the three-ratings-per item methodology" (p. 128).

Further, McLaughlin establishes that his finding [that generating three Angoff ratings represents no threat to the achievement levels setting process] cannot be easily dismissed as due to procedural differences:

"Although this study was not intended as an exact replication of the achievement-level setting process used for the NAEP, it is important to establish that it was sufficiently similar to warrant generalizations about the results" (1993a, p. 126).

CONCLUSION: Some evidence is presented in the NAE Report that judges exhibited some degree of inconsistency in generating item ratings (see a later section of this review). However, the issue of consistency is distinct from the issue of whether NAEP panelists were able to perform the Angoff rating task. 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, as illustrated by the references above. Further, the accumulation of research reports and project summaries of many educational and standard-setting entities details thousands of instances in which the Angoff method has been successfully utilized. Finally, speculation that the Angoff method requires an "impossible cognitive task" is refuted by studies commissioned by the NAE.

Objection 2 - The limitation of the Angoff procedure is not merely technical.

Because the central claim of the NAE Report is so clearly incorrect, it is logically unnecessary to consider claims that are based upon the refuted claim. However, the NAE Report does present an additional, different speculation that "the problems with the Angoff procedure are not merely technical" (p. 55). At the heart of this speculation is the claim that "by focussing on items one at a time, for example, the method prevents judges from arriving at an integrated conceptualization of what performance at each of the [achievement] levels should look like" (p. 55). Again, when the evidence upon which this claim is based is carefully scrutinized, the claim is not supported.

EVIDENCE: The preponderance of evidence from the NAE's commissioned studies suggests precisely the opposite conclusion about the opportunity for judges to form an integrated conceptualization of what performance at each of the achievement levels (Basic, Proficient, Advanced) should look like.

First, as one of the commissioned studies indicates, "the Board [i.e., NAGB] developed generic, but clearly distinct descriptors of the three achievement levels at each grade level as a first step to assuring clear distinctions in each subject area" (Pearson & DeStefano, 1993a, p.

7). Another NAE commissioned study details that participants in the NAEP levels-setting process did have the opportunity to arrive at integrated conceptualizations of student performance:

"After becoming familiar with the NAEP, the panelists first settled on an operational definition for basic, proficient, and advanced performance at their grade in their content area" (McLaughlin, 1993b, p. 77, emphasis added).

Incredibly, the NAE Report itself details that the NAEP levels-setting process was structured so as to ensure that levels-setting participants could arrive at integrated conceptualizations of student performance. The NAE Report details that: "panel members were given background materials to read prior to the meeting" (p. 30); that "panelists next took and scored a grade-appropriate NAEP booklet" (p. 30); that "panelists worked in small groups to generate lists describing what students should be able to do at each grade level to be considered basic, proficient or advanced" (p. 30, emphasis added); and that "after discussion to identify those elements that best exemplified each achievement level, panelists agreed on a final list of descriptors" (p. 31, emphasis added).

CONCLUSION: The assertion presented in the NAE report that "the [Angoff] method prevents judges from arriving at an integrated conceptualization of what performance at each of the [achievement] levels should look like" (p. 55) is wholly unsubstantiated. The evidence obtained from observation and documentation regarding how the process was conducted and the preponderance of evidence from the NAE's commissioned studies suggests precisely the opposite conclusion.

C) Conclusions Regarding Procedural Propriety and Results

The NAE Report presents several conclusions about the propriety of the procedures used to implement the Angoff procedure and conclusions about the reliability of the results of the levels-setting process. The major criticisms of the Report focus on the selection and training of "judges" (i.e., levels-setting participants) and concerns about the reliability of the ratings generated by the judges. In this section, I compare the procedures actually implemented in the levels-setting process with the procedures recommended in the relevant professional literature; I then evaluate these findings in light of the conclusions reached in the Report. Second, I evaluate the Report's conclusions related to reliability in light of the actual evidence that exists bearing on that question.

Selection and training of judges

Careful selection and proper training of judges to serve on standard setting panels is critical to the success of standard setting enterprises. These aspects of standard setting affect the likelihood that judges will reach consensus in providing their ratings and whether, as individuals, they will apply their judgment uniformly. The sampling procedures used to identify judges (i.e. the selection of judges from relevant populations) is directly related to whether consensus is necessary, desirable, or likely. The training used to familiarize judges with the methodology is related to the likelihood that judges will be able to consistently (i.e., reliably) apply their judgment. The following paragraphs address these issues of "interjudge consensus" and "intrajudge reliability."

Interjudge consensus

A pronounced criticism infusing the NAE Report is that Angoff judges failed to reach consensus regarding the achievement levels. Careful examination of this criticism reveals that it is only valid if consensus is defined as the agreement of all judges with one another, and if it is assumed that such consensus is both necessary and desirable. However, these assumptions

are critically deficient as preconditions for or evidence of a successful standard setting procedure. The deficiency stems from the fact that NAEP levels-setting panels were not constituted so as to contain like-minded persons representing a single, "approved" perspective. Such an approach would certainly maximize the likelihood that panelists would converge on a single, consensus standard. However, as experts in standard setting have noted:

"The judges [in standard setting] usually differ in life experiences, specialty, and professional skills. Often, in fact, the judges are selected specifically to provide a representative distribution of such background characteristics" (Plake, Melican, & Mills, 1991, p. 15, emphasis added).

Thus, congruent with NAGB policy and accepted psychometric practice, panels were constituted to represent a diverse and knowledgeable group of persons that could represent the varying and sometimes conflicting perspectives inherent in the pluralistic society to which the resulting standards would apply. This conscious effort toward the inclusion of diverse perspectives is documented in the NAE Report, which indicates that panelists were sampled from:

"individuals knowledgeable in each subject area...various professional and political organizations...teachers, nonteacher educators, and noneducators, and balanced with respect to gender, ethnicity, and geographic region" (NAE, 1993, p. 30).

The NAE Report verifies that the goal of a diversity was attained, noting that "the approach to selecting judges for 1992 ensured that the panels were broadly representative" (1993, p. 30).

However, an inclusive approach to constituting participants in the levels-setting process virtually assures that consensus--if defined as the agreement among all participants--will not be

attained. In fact, it is probable that the only way this kind of consensus can ever be achieved in standard setting is by fiat, or by using a sample of only one judge. But, such a definition of consensus is incongruent with both the context of a democratic society and the goals espoused by NAGB policy. It is also incongruent with current opinion within the psychometric community. An example from a recent journal article on standard setting illustrates that a sensitivity to differences, rather than an obsession with consensus is the more appropriate measure for judging standard setting procedures:

"Panelists in standard-setting studies should be chosen to represent all appropriate groups in the profession relevant to establishing the cutoff scores for the test. These panelists, therefore, will bring a diversity of knowledge, training, and opinions about the test and testing session to the rating session.... The goal [of training] is not to create an environment in which all judges agree on item difficulty, but rather to ensure that differences in ratings result from differences in perceptions of the item difficulty" (Mills, Melican, & Ahluwalia, 1991, p. 9).

In summary, the NAE Report raises questions about the "adequacy of the consensus process" (p. 52). The primary criticisms contained in the Report center on the findings that "there was great variation in individual judges' recommended cutscores around the group average;" that "individual variation did not diminish much by [the end of the process]" (p. 52); and that "feedback did not dramatically improve agreement in judgments" (p. 54). A review of the available data does reveal that there was variation between judges in the NAEP levels-setting process. That is, by the end of the process, there were differences between people regarding where appropriate achievement levels should be set.

In evaluating the observations of the NAE Report, one must note that it is certainly a matter of degree as to whether one terms a given amount of variation "great," whether a decrease in individual variation is "much" or whether improvement in agreement was "dramatic." These infinitely debatable matters of degree are not the deciding question, however,

and do not address the fundamental, underlying purpose for creating diverse panels in the first place.

What is critical is that the same data on variability can be interpreted to provide strong evidence for the success of the empaneling process. That is, the data may well reveal that the NAGB's intention to solicit broadly diverse input was successful. The evidence indicates that, over the rounds of ratings, the panelists moved toward consensus, although they never did reach perfect uniformity in their ratings. In the NAE Report, this fact is interpreted as indicating a weakness in the NAEP levels-setting approach. However, this phenomenon is entirely consistent with the expectations and reported findings in the psychometric literature, when standards are set by diverse groups of people representing varying perspectives. For example, in a study involving 236 judges using a modified Angoff methodology, Busch and Jaeger reported that:

"Small, but sometimes statistically significant, changes in mean recommended test standards were observed when judges were allowed to reconsider their initial recommendations following review of normative information and discussion" (1990, p. 145).

Additionally, the observation of small movement toward consensus is consistent with the design and goal of the procedure: that feedback provided to participants and their discussions of key issues served to refine and temper their judgements, but also that the desirable element of diverse perspectives was maintained throughout the procedure. As the Report concludes:

"within a range of p-values that individual judges were contemplating, participants were probably willing to select values that were still consistent with their own beliefs but moved in the desired direction as implied by feedback" (p. 50).

In conclusion, the Report's criticism that "the achievement levels in both reading and mathematics do not appear to reflect consensus standards or common expectations" (p. 54) is not compelling and is not necessarily even logically consistent with the a priori goals of the panel selection and training processes.

Intrajudge consistency

A second criticism in the NAE Report related to the selection and training of judges is that the judges were "internally inconsistent." That is, there is a concern that individual judges did not maintain personally consistent conceptualizations of performance at the three achievement levels. The NAE Report concludes that the extent of inconsistency is "unacceptable." As with conclusions regarding interjudge consensus, the evaluation of intrajudge consistency involves matters of degree. My review of the evidence presented in the NAE Report and companion studies leads to the conclusion that the evidence for internal consistency is mixed. On the one hand, the Report provides much positive evidence regarding consistency: "expected correlations between judges' estimated percentages and actual item p-values increased from [round to round] (p. 49);" "judges tended to rank-order items correctly" (p. 49); "information [provided to judges] served to improve consistency correlations" (p. 49); and "the final levels in reading and mathematics follow a reasonable pattern across grades" (p. 50).

On the other hand, studies commissioned by the NAE also illustrate sources of inconsistency within the panelists. For example, participants were inconsistent in applying a uniform standard to multiple-choice and extended response items and for easy versus hard items (see McLaughlin, 1993b).

The NAE Report synthesizes this conflicting evidence and concludes that "the results did not suggest any systematic variation in recommended cutpoints due to substantive dimensions of the assessment" (p. 48). Although there are many reasonable conclusions and proposals that could be made based upon the evidence, the NAE Report's conclusion is not one of them. In my opinion, the conflicting evidence is just that--conflicting evidence. Such evidence probably calls for: 1) further investigation of the sources of information judges actually utilize in

generating their ratings; and 2) inclusion of specific directions, examples, monitoring, and feedback to panelists regarding the influence of item format and difficulty during standard setting training and actual implementation of the procedure. However, the suggestion that inconsistencies due to item format and difficulty invalidate the resulting levels is inappropriate and inconsistent with current notions of validity as a matter of degree (see Messick, 1988).

D) Conclusions Regarding the Validity of the NAEP Levels

The opinion expressed in the NAE Report is that the NAEP achievement levels lack validity. That opinion derives from three beliefs: 1) that faithful implementation of a professionally accepted procedure does not provide evidence of validity; 2) that achievement levels were set "too high;" and 3) that the levels are not sufficiently related to the NAEP frameworks. In the following paragraphs, I evaluate the arguments supporting these three beliefs, and draw conclusions about the validity of the NAEP achievement levels.

1) Procedural evidence for validity

As has been described in the NAE Report and demonstrated in this review, NAGB implemented a process to accomplish the goal of gathering and summarizing opinion regarding what the achievement levels should be. And, the Report recognizes that "a standard setting method, even when implemented precisely as recommended in the literature, will not necessarily produce a true or valid standard. It will only summarize the judges' opinions" (p. 23).

It is essential to perceive the manner in which the issue of validity is cast within the NAE Report:

- 1) The Report states that a carefully followed procedure does not "guarantee" a valid result. This is, of course, strictly true, though perhaps fatuous. Careful attention to procedure only guarantees that a procedure was carefully followed. However, it is most definitely true that fidelity to carefully prescribed, professionally accepted procedures provides evidence of validity. By failing to

recognize that procedural fidelity is an important element in evaluating the validity of the achievement levels setting process, the Report makes a critical omission.

Such an omission is not inconsequential. It means that the NAE Panel failed to even consider the relative weight of procedural fidelity in evaluating the overall validity of the process and resulting standards. Unfortunately, therefore, the NAE Report was, by design, incapable of synthesizing and evaluating this primary source of validity evidence. Thus, even if--as much of the technical documentation indicates--the procedure was technically appropriate, based upon broad input, founded on professionally acceptable principles, and carried out exactly as prescribed, the NAE evaluation would fail to capture this information.

2) The NAE Report indicates that the process used does not necessarily lead to a "true" or "valid" result. This is clearly incorrect. As cited in the NAE Report, the result of the Angoff standard setting procedure is a summarization of the particular panelists' judgment regarding appropriate achievement levels. In this regard then, a "true" result is one that accurately summarizes the judgments. If properly sampled from a population of suitable persons, and if the Angoff procedure was followed carefully (see above) then the participants' judgments regarding achievement levels are the best estimates of where the population of similarly qualified judges would set the achievement levels.

Again, appropriate sampling and procedural fidelity provide strong evidence for whether the process resulted in "true" or "valid" achievement levels.

In summary, in evaluating the validity of the NAEP achievement levels, the NAE Report makes a critical error in failing to investigate and evaluate available evidence of procedural

validity. The Report fails to recognize that the contributions of independent judges, sampled from a diverse population, properly trained, and participating in a professionally accepted process, add substantial support to the goal of establishing credible, valid standards of performance.

2) How high is too high?

Another prominent contention of the NAE Report is that the NAEP achievement levels are "too high." The Report states that: "The weight of evidence suggests that the 1992 achievement levels were set unreasonably high" (p. xii). In another place, the Report calls the levels "unrealistically high" (p. 28). However, to evaluate these judgments, it is essential to review appropriate background for the NAEP levels-setting process.

First, as the Report indicates, the NAEP levels-setting process was initiated in response to a Congressional directive which charged NAGB with "identifying appropriate achievement goals for each age and grade in each subject area to be tested under the National Assessment" (p. viii, emphasis added). Thus, the charge was given to establish desirable performance levels, not simply to--in my opinion--waste time and resources validating a status quo. (A test would be unnecessary if that is the motivation for giving one.) If, as the Report indicates, a large number of students do not achieve the level of expectation implicit in the performance levels, then that unfortunate result may not be so much a matter of unrealistic expectations as unrealized potential (for whatever reason).

Again, it is important to realize that NAGB was charged with establishing aspirations, as opposed to (or in addition to) reflections, of educational progress. The salience of this point is highlighted in the NAE Report as the "should" versus "can" issue (p. 89). In short, the levels-setting process was designed to elicit and summarize the opinions of a diverse and representative group of persons regarding what level of performance should be expected of American students at grades 4, 8, and 12. As the Report indicates, "achievement levels...are intended to establish expectations for what students should know and be able to attain at each level. In this sense the standards are statements of desired rather than actual outcomes" (NAE, 1993, p. 89, emphasis added).

Logically, then, if high standards result from the levels-setting process, it cannot be said that the standards are "too high." The standards are merely as high as the participants' aspirations for students' performance. It seems somewhat improper for the NAE evaluation to refute the aspirations of those empaneled to establish achievement goals. Interestingly, the NAE Report acknowledges this fact: "Judges are expected to envision performances that should be; therefore, their judgments do not necessarily have to conform to normative data" (p. 43, emphasis in original). Thus, in essence, claims that the performance goals set by the levels-setting panels are "too high" are simply invalid on their face.

Second, despite the fact that the "too high" claim cannot be supported at its most fundamental level, evidence presented in the NAE Report regarding this issue can be subjected to examination. A review of the evidence commissioned for the NAE Report does not provide support for the contention that the levels are "too high;" rather, a good deal of evidence that supports the validity of the achievement levels is found in the NAE studies. It is unclear why the NAE Report did not weigh the evidence in making its determination.

As the NAE Report acknowledges, "no external criteria can serve as the ultimate authority in judging what constitutes advanced, proficient, or basic performance" (p. 15). Nonetheless, the NAE commissioned studies to relate performance on the NAEP to other indicators of domestic and international student performance. Although available for the Report, the following evidence supporting the validity of the achievement levels was overlooked, ignored, incorrectly weighted, or not reported:

- Results of the reading and mathematics validation panels. These results are not reported in the NAE document (see p. 32, NAE Report);
- Results of "Researcher Validation of Teacher Ratings (see p. 65, NAE Report). These results, which ostensibly threaten the validity of the achievement levels, are based upon a methodologically silly investigation. In gathering data for the study, "teachers were urged to pay attention to the definitions and not to rate children

in relation to others in the class" (p. 65). In contrast to the detailed materials, rigorous training, and monitoring received by the actual achievement levels panels, "urging" teachers to read and follow a set of written directions is laughable as a "replication" or "validation" study methodology. In addition, researchers and teachers disagreed in classifying students as below basic, basic, proficient and advanced more than 50 percent of the time--hardly a "validation" (McLaughlin, 1993d, pp. B-6, B-7).

- Results of "Contrasting Groups Studies" (see p. 65, NAE Report) are based upon incomparable training in applying the achievement levels descriptors. As noted in the NAE Report, variations in training can dramatically affect resulting standards. In this instance, the training provided was so markedly different that comparisons are nearly meaningless. In addition, there are significant methodological deficiencies in this study. The most fundamental is that NAEP scores were computed for individual students, even though it is well-known that NAEP is not designed to give individual scores and that such scores are unreliable.

- Results of the "Kentucky Comparisons" (see p. 74, NAE Report) are markedly underemphasized. From a research design perspective, the Kentucky study represents a highly relevant comparison. As noted in the NAE Report, "Although many states have assessment programs that would permit indirect comparisons with NAEP results, Kentucky is rare in that it uses achievement levels similar to the NAEP levels" (p. 74). The highly pertinent finding that the NAEP and Kentucky assessments "produced similar percentages of students in the upper two categories" (p. 74) and that the Kentucky studies "support the reasonableness of the NAEP eighth-grade proficient and advance cutpoints in mathematics" (p. 75) would have received much greater weight in an evenhanded evaluation.

- Results from the "Content Expert" studies in reading and mathematics (see p. 78, NAE Report) are underemphasized. In these instances, the NAE commissioned studies found that, for example, "content experts in reading consistently identified cutpoints above the official achievement-level cutpoints" (p. 78, emphasis added). An accurate interpretation of these findings would be that the NAEP achievement levels were actually more reasonable than those set by the content experts. In an incredible statement reflecting a departure from scientific objectivity, the NAE Report states that "the [NAE] Panel did not in fact take the final [content expert study] results as either confirming or disconfirming of the NAGB levels.

- Results of the NAE commissioned "Rated Achievement Levels of Completed NAEP Mathematics Booklets" (McLaughlin, 1993c) are not reported. In this case, the NAE Report ignores the results of one of its own validation studies which found that "the official cutpoints of 256, 294, and 331 [for Basic, Proficient, and Advanced achievement levels] are not substantially in conflict with the panelists' ratings in this study" (p. 11, emphasis added). Interestingly, the same study also shed light on the NAE Report's claim that the Angoff procedure as implemented in the achievement levels setting process resulted in an unacceptable degree of interjudge variability (previously discussed in this review). On this matter, the NAE commissioned study found that "the variation in results among panelists in this study was similar to that observed in St. Louis, although the St. Louis sample was more broadly representative of the nation" (p. 11).

Finally, a third perspective on whether the NAEP achievement levels were set "too high" can be found in abundance in the literature of educational reform. For example, the NAE Report cites the well-known A Nation at Risk (U.S. Department of Education, 1983) in which the "rising tide of mediocrity" and generally woeful state of American educational achievement

are described. The Nation at Risk report was followed by numerous other reports that articulated the same message: American students were not performing as well as they should perform. When combined with other evidence (e.g., that SAT scores have not increased during the period since the Nation at Risk report was issued; that teachers' judgments of student performance are often inaccurate, that teaching methods, etc., have not changed substantially in the interim, and so forth) the finding that unacceptable percentages of students reach the Basic, Proficient, and Advanced levels is unsurprising. Indeed, the fact that application of the NAEP achievement levels reveals results that are strikingly consistent with widespread perceptions of American educational achievement--and confirmed by content experts--provides strong evidence for the validity of the achievement levels.

3) Relationship to the NAEP Frameworks

A third criticism of the NAEP levels-setting process contained in the NAE Report is that the judgments of levels-setting panelists were not sufficiently related to the NAEP frameworks (see p. 40ff, NAE Report). However, there is little high-quality evidence to support this notion.

First, there is a serious methodological flaw in the primary NAE-commissioned study upon which the Report's conclusion is based (see Pearson & DeStefano, 1993). Specifically, the Pearson and DeStefano study reported that: "when participants who set the [reading] achievement levels were asked in a mail survey to describe a good reader 2 weeks after the level-setting session, they gave a great variety of answers" (NAE, 1993, p. 40). Unfortunately, the methodological choice of a mail survey sent to participants two weeks after the levels setting process had ended did not even have the potential to answer the question of interest: that is, how much did participants rely on the reading framework? It is important to note that during the two week interval, training, discussions, and monitoring did not occur. The tendency to drift from appropriate referents is the very purpose for conducting standard setting meetings in a single location with intensive training, discussion, monitoring, etc.,. It is unsurprising that, when these controls were removed, panelists' ability to retain appropriate conceptualizations decayed. Thus, the evidence cited in the NAE Report does not directly bear on the question of interest.

A second primary issue raised in the NAE Report is the "initial failure to tie achievement levels to NAEP frameworks" (p. 97). The Report concludes that "the initial achievement-level descriptions were not adequately connected to the framework but relied more on judges' personal experience" (p. 97). The Report presents an "ideal" model of relationships among components in the achievement levels setting process (see Figure 3.2., NAE Report). This ideal model is contrasted with an "actual" model of relationships in the levels-setting process (see Figure 3.3, NAE Report). Ostensibly, because the NAEP levels-setting process "violates" the ideal model, then the validity of the achievement levels is threatened.

Two aspects of this modeling are noteworthy. First, the ideal model is not an accepted model of validity; it is important that this "ideal" model is not referenced to any empirical or theoretical work in the field of standard setting. In fact, the components in the "actual" model are probably a better model of the experience-based, item-based, and social influences on the standard setting process (see Fitzpatrick, 1989; Norcini, Shea, & Kanya, 1988; Smith & Smith, 1988). The components of the "actual" model account for the influences of feedback to judges, personal experience, and diversity of opinion. The "ideal" model denies the importance of these influences. In summary, on theoretical grounds, the model of validity held as the criterion in the NAE Report is actually poorly specified. The model of influences as actually observed is closer to reality.

A second aspect of the NAE conclusion that "the initial achievement-level descriptions were not adequately connected to the framework but relied more on judges' personal experience" warrants closer scrutiny. A review of the studies commissioned by the NAE does reveal anecdotal evidence that some of the panelists considered their own students, their own experiences, their own knowledge of reading, etc., as they participated in the levels-setting process. However, two considerations are noteworthy. First, as mentioned above, it should be expected that personal experience, professional knowledge, etc., would infuse the entire levels-setting process. These kinds of relevant knowledge, experiences, etc., are the very reason that the particular group of panelists was selected. Second, and important from a critical perspective on the research supporting the Report's conclusion, is that no evidence was gathered to support

the claim that panelists relied more on their personal frameworks than on the content frameworks and descriptors. It is indeed a leap to suggest that because panelists called upon their relevant knowledge and experience, that they relied to a greater degree on these aspects than on the frameworks. The Report's suggestion that an "either/or" phenomenon occurred, i.e., that "Panelists...used personal experience and opinions to develop the descriptions and make item judgements rather than following the framework" (p. xii), is not psychologically realistic or supported by any of the available evidence presented in the NAE commissioned studies.

II. ACCURATE CONCLUSIONS OF THE REPORT

Several conclusions and recommendations of the NAE Report are accurate and merit further consideration. In this section of my review, I identify and explain some of the most important positive contributions of the Report.

A) Link Between Achievement Level Descriptions and Angoff Ratings

The NAE Report concludes that "the process for developing the descriptions in reading and mathematics was inadequate because it did not ensure that final descriptions were agreed upon before attempting to set cutscores" (p. 60). The Report and the companion studies provide unambiguous evidence that achievement levels panelists based their item ratings on one set of descriptions; the finalized descriptions differed to some degree from the ones used by the panelists. This is poor practice. It is reasonable to assert, as the NAE Report does, that consistency of meaning between descriptors and achievement levels is essential for maximizing the interpretability of NAEP scores. To ensure that valid interpretations could be made, no changes in the descriptors should have occurred.

On the other hand, the effects of these changes are unknown and not clearly illuminated by the NAE commissioned studies. One of the NAE commissioned studies provided a direct comparison of achievement levels set using varied descriptors:

"This study also offered an opportunity for two additional comparisons. First, the comparison of cutpoints generated between the original (St. Louis) version of the achievement-level descriptions and the revised (Nantucket) descriptions indicated no significant differences at the basic and advanced levels and only a marginally significantly higher proficient cutpoint by panelists using the Nantucket descriptions. We do not believe that the level of discrepancy observed in this study is cause, in itself, for judging the revisions made in Nantucket to have invalidated the achievement levels" (McLaughlin, 1993a, p. 128, emphasis added).

In short, the McLaughlin study revealed that the revision of descriptors had little effect. The only statistically significant difference was for a version of the descriptions that would have resulted in an even higher standard.

Despite the availability of some evidence that the revision of descriptions was inconsequential, the "true" effect cannot be stated with certainty. The NAE Report reflects this uncertainty by employing the "may" argument: The Report concludes that "the final descriptions may not be valid for describing the assessment and may not correspond to the cutpoints determined on the basis of earlier definitions" (p. 60, emphasis added). The "may" argument should be interpreted in two ways. By one interpretation--the interpretation suggested in the Report, the final descriptors may not be valid for describing the assessment; on the other hand, another interpretation is that they may be perfectly valid for describing the assessment.

In another place, the Report states that the cutpoints are "potentially inconsistent with the narrative descriptions" (p. 58). Again, this also means that the cutpoints are also potentially consistent. The point is that no conclusive evidence has been gathered to settle the issue of consistency or equivalence of the descriptors. It is unclear what effect the revision of descriptors had, although the available evidence indicates that the effect of the revisions would be slight. The Report clearly errs in concluding that revision of the achievement levels descriptors "invalidates the agreed upon cutpoints" (p. 57). A more accurate conclusion would be that any

revision of the achievement levels descriptors is likely to have had some unknown degree of effect.

The inaccurate conclusion presented in the Report is particularly striking given the evidence that was available to the NAE Panel. For example, in describing the investigations that were commissioned to examine the degree of effect, the NAE Report indicates that "the results of these experimental comparisons were inconclusive" (p. 58). In contrast to the strongly-worded conclusion that the revision "invalidates the agreed upon cutpoints" is the more reasonable conclusion--in the same Report--that "the Panel could not draw any general conclusion about the magnitude of changes likely to occur in cutpoints when substantive descriptions are changed" (p. 58). The latter conclusion is the only one supported by the available evidence. It is unclear why the Report highlights the much stronger and less supportable interpretation.

B) Within-Grade Score Reporting

The recommendation to implement within-grade score reporting does not derive directly from any of the work commissioned by the NAE. However, many other researchers have produced extensive discussions of the problem of a single vertical scale for describing educational progress. It is a reasonable summary of the literature that a satisfactory scaling of complex subject matter, involving developed abilities over several years, is extremely difficult to attain; current attempts to do so are plagued with problems of interpretability. The NAE recommendation to move toward within-grade scales seems to be a conservative and prudent recommendation in the absence of evidence documenting the superiority of a single cross-grade scale.

C) Weaknesses in NAEP Item Pools

At several points, and in varying contexts within the NAE Report, the issue of weakness in the NAEP item pools arises. It is observed that there were not enough items representing higher level performance, especially at the Advanced level. Similarly, the Report notes that the

quality of released NAEP items is poor. Because I have not reviewed all of the evidence upon which these observations are based, I cannot offer a confident evaluation of this criticism. However, throughout the course of the Report and the companion studies, the issue was raised in enough varied and independent contexts that the matter clearly deserves additional attention.

On the other hand, the issue of weaknesses in the NAEP item pools does not seem to bear directly on the achievement levels setting process. It is not shown in the NAE Report or demonstrated in the companion studies that suspected weaknesses adversely affected the NAEP achievement levels.

D) Cutpoint Adjustments

The NAE Report asserts that a concern regarding the final achievement levels is the adjustment (downward) of the mathematics achievement levels by one standard error. The report is correct in observing that "there is ample justification in the technical literature for revising the proposed cutpoints in response to additional information" (p. 56). However, at least one of the objections in the Report deserves consideration. Specifically, no compelling justification is provided for why the mathematics cutpoints were adjusted and the reading cutpoints were not. If there is "additional information" that justifies adjusting a set of cutpoints, the same information related to the other sets of cutpoints should be scrutinized. Further, a consistent rule for adjusting cutpoints must be applied, and any additional information used to adjust cutpoints should be evaluated according to the same criteria. Any adjustment should be made based on rational, explicit, clearly articulated, criteria (see Geisinger, 1991).

In summary, the NAE Report contains some accurate conclusions that deserve further consideration. Four such conclusions are listed above. In reviewing these conclusions, however, it is interesting to note that only two of the four are related to the achievement levels setting process (i.e., revisions in frameworks, and cutpoint adjustment). The other issues bear more directly on the construction of the NAEP itself (i.e., item pool weaknesses and scaling).

III. ANALYSIS OF NAE REPORT RECOMMENDATIONS

The NAE Report goes beyond an appraisal of the process used by NAGB for establishing achievement levels for the NAEP and, in many instances, recommends alternatives. Logically, if sound alternatives exist to the process utilized by NAGB, criticism of the NAGB approach would be warranted. Conversely, if sound alternatives cannot be suggested, the NAGB approach would be supported. In this section, I identify three such recommendations suggested in the NAE report and evaluate their potential for yielding defensible achievement levels.

A) Contrasting Groups Methodology

The NAE Report recommends that, for the future, "[achievement] levels be validated using the contrasting groups method that was used to evaluate the current set of achievement levels" and that "performances both within and across grade should be examined to ensure that the observed patterns confirm predictions based upon the explicit or implicit conceptual models used to formulate the performance standards" (p. xix). These recommendations should be viewed skeptically for two reasons.

First, as described in the Report, "to implement the contrasting-groups design, judges or raters must have knowledge of what students can do in the domain measured by the test,...must be able to apply the definitions of advanced, proficient, and basic performance, and to classify students into appropriate categories" (p. 64). This is an accurate description of the contrasting groups methodology. However, note that the sophisticated knowledge, difficult classifications, and complex conceptualizations that apply to judges using the contrasting groups procedure are nearly identical to the characteristics of the Angoff procedure. The two methodologies address a common goal, but rely on similar psychological aspects--aspects which the Report objected to as "impossible cognitive tasks." If it were true, the same objection related to the difficulty of the cognitive task in an item-judgment approach would also apply to the contrasting groups design.

Although it is my opinion that the contrasting groups design is somewhat more easily implemented than the Angoff approach, the contrasting groups design still depends on the ability of humans to apply their judgment in difficult, often high-stakes circumstances. Contrary to the implication of the NAE Report, the contrasting groups design does not represent the magic bullet of standard setting, and for other reasons may not be the methodology of choice for setting achievement goals.

B) Linkage to Emerging National Standards

In its rejection of the achievement levels established for NAEP, the NAE Report is critical of nearly every aspect of the NAEP. It asserts that essentially everything about the National Assessment of Educational Progress is invalid--test development procedures, item pools, test content, achievement levels, reporting strategies, and so on.

In particular, the Report is critical of the NAEP because "current NAEP item pools...are not sufficiently congruent with emerging national content standards" (p. xiii) and because the NAEP achievement level descriptions "cannot adequately represent ideal future-oriented standards" (p. xiii). In another place the Report recommends that the NAEP should "provide a stable basis for comparison as well as for evolutionary change" (p. xx).

Even a casual reading of these criticisms reveals their logical silliness. Nonetheless, a critical examination of these recommendations is in order as a means of illustrating that the suggested alternatives would represent a serious weakening of the current achievement levels.

First, one must ask the obvious questions about the Report's recommendations: How can standards be anchored to "emerging" content? How can any assessment ever be linked to "ideal future-oriented" content or standards? How are the goals of providing stable bases for comparison as well as evolutionary change to be achieved? I don't think these things make any sense. I suspect that the Report provides so little detail about these recommendations because they cannot be explained. As coherent statements about the relationship between curriculum and assessment they are illogical. No assessment should be expected to hit a moving invisible target. As recommendations for NAEP policy they are seriously deficient.

However, beyond a surface appraisal of their logic, a deeper examination actually yields strong support for the current achievement levels-setting process. At a general level, it is important to note that the National Assessment should reflect what is currently taught in American classrooms. If the NAEP were linked to "emerging" or nascent content, it would actually weaken the test as a measure of what American students know and can do. That is, to follow the NAE recommendation would seriously threaten the validity of the NAEP for its intended uses.

A closer look at the evidence gathered as part of the NAE evaluation reveals that, as currently constituted, the NAEP achievement levels are a considerably better match with the curriculum actually experienced by American students than would be the match with "emerging" content. For example, the NAE Report indicates that "experts also criticized the quality of items, finding that as a whole they did not adequately reflect some of the NCTM [National Council of Teachers of Mathematics] content standards" (p. 84, emphasis added). It is instructive to evaluate what this criticism actually means.

First, the Report reveals that the NAEP items do not adequately reflect some of the NCTM standards. This necessarily means that the NAEP items do reflect many (or most?) of the standards.

Second, is it an error that the NAEP items do not reflect all of the NCTM standards? To argue that the NAEP should be entirely consistent with all of the NCTM standards, places a severe burden on critics to demonstrate that the NCTM standards are actually in place, and an integral part of the classroom experiences of American school children. Unfortunately, it is well known that the NCTM standards are not completely in place, are not as widely or as thoroughly implemented as many would hope, and do not yet represent the status quo in American education. As the NAE Report indicates: "The NCTM standards have been extolled for 3 years and still have not reached most classrooms" (p. 113).

Thus, while it is true that the NCTM standards are gradually increasing in prominence and in practice in American classrooms, they represent an invalid basis for current assessments.

Sound testing principles dictate that a defensible assessment would evolve to reflect that increase. Current NAGB policy, cited in the NAE Report, mirrors these sound testing principles:

"incorporation of [emerging] standards into the National Assessment should be done through successive adjustments of its frameworks and assessments and that the goal should be to achieve a balance between the vision contained in new voluntary standards and the reality of current instruction" (p. 114).

In this light, the observation in the NAE Report that as a whole, the NAEP items do not yet reflect all of the NCTM content standards is, on balance, a good thing. As has been noted: "NAEP assessments are designed by broad-based consensus groups and emphasize material that is commonly taught for each grade and subject tested" (USGAO, 1993, p. 9). Thus, although the Report interprets this finding in a negative light, it actually reflects well on the NAEP standards and supports their validity.

Finally, one should consider what might occur if, outpacing the curricula and experiences of American students, the NAEP were to reflect all of the NCTM standards. Undoubtedly--and justifiably--the NAEP would be subjected to charges of "psychometric imperialism" (i.e., of testing experts imposing their ideas on American education)(Madaus, 1988). Currently, the nearly unanimous position of the measurement community is clear: "Tests should be for monitoring but should not drive instruction" (Shepard, 1991, p. 5). In summary, the recommendation to use "emerging" standards as a criterion by which the appropriateness of the current NAEP and NAEP achievement levels would, if implemented, seriously harm the credibility and validity of the National Assessment.

C) Usage of Percentiles as Standards

A final recommendation of the NAE Report is that, as an alternative to the NAEP achievement levels, standards of performance:

"could be set at three levels using thoughtfully chosen percentile scores... The Panel suggests the 95th, 75th, and 25th percentiles for a base year could be used as benchmarks against which to measure future progress" (p. xvi).

This recommendation, if implemented, has the potential to seriously undermine the credibility of the NAEP program. First, this recommendation is contrary to the warnings contained in the Standards for Educational and Psychological Measurement (AERA/APA/NCME, 1985) and in measurement textbooks. For example, one textbook advises that, "Perhaps the greatest mistake is to interpret norms as standards" (Mehrens & Lehmann, 1984, p. 316). Contrary to this advise, the Report recommends that future performance standards should be established at arbitrary levels based upon current norms.

Additionally, the recommendation is particularly ill-advised given the recent history of misuse of norm-based scores in American education. One need only recall the national furor over the "Lake Wobegon Report (Cannell, 1989). In that case, a widely discussed report illustrated the misuse of norms and described "How Public Educators Cheat on Standardized Achievement Tests." The resulting loss of public confidence in the results of large scale tests will take much effort to restore. If implemented, the NAE Report's recommendation could easily witness a national assessment system in which, for example, 50% of American students could be performing above the 95th percentile. Given the apparent imprudence of repeating the "Lake Wobegon" fiasco, the NAEP achievement levels represent a far better alternative.

CONCLUSIONS AND OBSERVATIONS

The NAE Report on the NAEP achievement levels-setting process utilized by NAGB in 1992 contains some positive evaluation and recommendations, as well as negative judgments about the process utilized by NAGB in setting performance levels. However, in my opinion, the Report provides an overwhelmingly and overly negative description of the NAEP levels-setting process--a view that is not supported by evidence available for the NAE Report.

In summarizing the results of my review, it is my opinion that the conclusions of the NAE Report: 1) rely on the input of researchers who do not possess relevant expertise in the area of standard setting; 2) do not derive from the application of accepted evaluation guidelines, criteria, or procedures; 3) are presented in a systematically unbalanced manner; 4) are based upon research studies that were not particularly well-suited to answering the questions of interest; and 5) lead to recommendations that would substantially harm the credibility and validity of the National Assessment of Educational Progress.

However, despite the identification of these serious flaws, it should not be concluded from the above evaluation that the NAE Report is without merit. The NAE Report identified issues associated with the levels setting process that warrant further investigation, and issues related to NAEP item development and scaling that are problematic. It can be said that the levels- setting process is not without residual difficulties and drawbacks. On the contrary, because the nature of all standard setting is judgmental, all standard-establishing procedures can be refined and improved. It is unlikely that any process could be designed and implemented in such a way as to be beyond reproach. As Jaeger has observed:

"If the literature on standard setting is conclusive on any point, it is the difficulty of setting defensible standards on competency tests. There is no agreement on a best method, although some procedures are far more popular than others" (1988, p. 491).

Further, as Livingston and Zieky (1982, p. 61) have candidly observed: "You will never be able to prove that your passing score is correct." And, just as a passing score cannot be proved correct, a test or levels-setting procedure cannot be "proved" to be valid--or invalid. As Messick has noted:

"What is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators... It is important to note

that validity is a matter of degree, not all or none... [It] is an evolving property and validation is a continuing process" (p. 13, emphasis added).

The accumulation of evidence from all sources--National Assessment Governing Board, National Academy of Education, American College Testing, etc.--must be weighed to arrive at any conclusion as to whether the NAEP achievement levels contribute to accurate inferences regarding student performance.

In the process of gathering some of that evidence, the NAE Report has identified some possible procedural and technical refinements for the future. As appropriate, these contributions are discussed in Section II of this review. However, by accepted standards of psychometric propriety, and when evaluating all of the available evidence, it must be concluded that the process utilized in establishing the NAEP performance levels was executed overall in a psychometrically sound manner. Further, the process resulted in the establishment of achievement levels that can be defended as faithfully responding to the Congressional directive to identify appropriate achievement goals for the ages, grades, and subjects covered by the NAEP. This success is clearly not reflected in the NAE Report.

Finally, it is my observation that the NAE Report is not a particularly adequate document to inform discussions of the quality of the NAEP achievement levels-setting process for two reasons. First, in reading the Report one is struck by the fact that it is more relevant to discussions of reshaping the National Assessment program as a whole. As such, it is not a technical evaluation of the levels-setting process, but a distinctly policy-oriented document. And, the Report does not hide this orientation. For example, in one place the Report notes that "Two themes are critically important in defining the policy context for this evaluation" (p. xi, emphasis added). In another place, Report candidly reveals that "the fundamental issues addressed in this report are not technical ones" (p. 13). Clearly, it is not relevant to my review to judge whether the Report's broad NAEP policy-setting orientation is appropriate within the context of an evaluation of the achievement levels-setting process. However, it is clear that such

a broad policy-setting orientation is beyond the Congressional directive to the National Assessment Governing Board. In my opinion, that charge has been adequately carried out.

Second, the critical stance of the NAE Report toward the establishment of achievement levels for the NAEP represents an extension of a critical stance on the part of the NAE toward competency standards generally. In a 1978 report the NAE argued strongly against then-current reform efforts as well--using the same rhetoric as the present Report:

"The NAE Panel believes that any setting of state-wide minimum competency standards--however understandable the public clamor which has produced the current movement and expectation--is basically unworkable, exceeds that present measurement arts of the teaching profession, and will certainly create more social problems than it can conceivably solve" (NAE, 1978, p. iv).

Although the earlier NAE Report should be credited for bringing attention to potential concerns, it is clear that much of the improvement in American students' basic skills would have been missed had the Report's advice been heeded. In a similar vein, though caution is appropriate, the present NAE Report has the potential to weaken a similar impetus for educational reform.

REFERENCES

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (pp. 508-600). Washington, DC: American Council on Education.
- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.
- Busch, J.C., & Jaeger, R.M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. Journal of Educational Measurement, 27(2), 145-163.
- Cannell, J.J. (1989). The "Lake Wobegon" report: How public educators cheat on standardized achievement tests. Albuquerque, NM: Friends for Education.
- Colton, D. A. & Hecht, J. T. (1981, April). A preliminary report on a study of three techniques for setting minimum passing scores. Symposium presentation at the annual meeting of the National Council on Measurement in Education, Los Angeles, CA.
- Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-129.
- Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. Review of Educational Research, 59(2), 315-328.
- Geisinger, K.F. (1991). Using standard-setting data to establish cutoff scores. Educational Measurement: Issues and Practice, 10(2), 17-22.
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational measurement, 3rd ed. (pp. 485-514). New York: Macmillan.
- Jaeger, R.M. (1991). Selection of judges for standard-setting. Educational Measurement: Issues and Practice, 10(2), 3-6, 10, 14.
- Joint Committee on Testing Practices (1988). Code of fair testing practices in education. Washington, DC: Author.

Klein, L. W. (1984, April). Practical considerations in the design of standard setting studies in health occupations. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Livingston, S.A., & Zieky, M.J. (1982). Passing scores. Princeton, NJ: Educational Testing Service.

Madaus, G.F. (1988). The influence of testing on the curriculum. In L.N. Tanner (Ed.), Critical issues in curriculum: Eighty-seventh yearbook of the National Society for the Study of Education (pp. 83-121). Chicago: University of Chicago Press.

McLaughlin, D.H. (1993a). Order of Angoff ratings in setting multiple simultaneous standards. Report prepared for the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment Project.

McLaughlin, D.H. (1993b). Validity of the 1992 NAEP achievement-level-setting process. Report prepared for the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment Project.

McLaughlin, D.H. (1993c). Rated achievement levels of completed NAEP mathematics booklets. Report prepared for the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment Project.

McLaughlin, D.H., et. al. (1993d). Teachers' and Researchers' Ratings of Student Performance and NAEP Mathematics and Reading Achievement Levels. Report prepared for the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment Project.

Mehrens, W. A., & Lehmann, I.J. (1984). Measurement and evaluation in education and psychology. New York: Holt, Rinehart and Winston.

Meskauskas, J. A. (1986). Setting standards for credentialing examinations: An update. Evaluation and the Health Professions, 9, 187-203.

Messick, S. (1988). Validity. In R.L. Linn, Educational measurement, third edition (pp. 13-103). Washington, DC: American Council on Education.

Mills, C. N. & Melican, G. J. (1988). Estimating and adjusting cutoff scores: Features of selected methods. Applied Measurement in Education, 1, 261-275.

Mills, C.N., Melican, G.J., & Ahluwalia, N.T. (1991). Defining minimal competence. Educational Measurement: Issues and Practice, 10(2), 7-10.

National Academy of Education (1978). Improving educational achievement. Washington, DC: United States Government Printing Office.

National Academy of Education (1993). Setting performance standards for student achievement. Washington, DC: Author.

National Assessment Governing Board (1991). NAGB Policy Framework and Technical Procedures for Setting Appropriate Achievement Levels for the National Assessment of Educational Progress. Washington, DC: Author.

Norcini, J.J., Shea, J.A., & Kanya, D.T. (1988). The effect of various factors on standard setting. Journal of Educational Measurement, 25(1), 57-65.

Pearson, D. & DeStefano, L. (1993a). An evaluation of the 1992 NAEP reading achievement levels; Report one: A commentary on the process. Report prepared for the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment Project.

Plake, B.S., Melican, G.J., & Mills, C.N. (1991). Factors influencing intrajudge consistency during standard-setting. Educational Measurement: Issues and Practice, 10(2), 15-16, 22, 25.

Rock, D. A., Davis, E. L., & Werts, C. (1980, June). An empirical comparison of judgmental approaches to standard-setting procedures (ETS Research Report). Princeton, NJ: Educational Testing Service.

Shepard, L. (1991). Psychometricians' beliefs about learning. Educational Researcher, 20(7), 2-16.

Smith, R. L. & Smith, J. K. (1988). Differential use of item information by judges using the Angoff and Nedelsky procedures. Journal of Educational Measurement, 25, 259-274.

United States Department of Education (1983). A nation at risk: The imperative for educational reform. (Washington, DC: Author).

United States General Accounting Office (1993). Educational achievement standards: NAGB's approach yields misleading interpretations (Report No. GAO/PEMD-93-12). Washington, DC: Author.

Comments on the NAE Evaluation of the NAGB Achievement Levels

Michael Kane

**Professor
Department of Kinesiology
University of Wisconsin, Madison
September 1993**

Section 1 - General Comments

The National Academy of Education (NAE) Panel has drawn two major conclusions in its evaluation of the National Assessment Governing Board's (NAGB's) efforts to set achievement levels for the National Assessment of Educational Progress (NAEP). In its summary report, *Setting Performance Standards for Student Achievement, A Report of the National Academy of Education Panel of the Evaluation of the NAEP Trial State Assessment: an Evaluation of the 1992 Achievement Levels*, the NAE Panel concludes that "... the Angoff procedure is fundamentally flawed for the setting of achievement levels" (p. xii), and that "The weight of evidence suggests that the 1992 achievement levels were set unreasonably high" (p. xii). However, the evidence presented in the final report of the NAE Panel and in the reports of the studies it commissioned do not justify these conclusions and some of the evidence directly contradicts the conclusions.

The NAE report is well written and provides a very insightful discussion of the general issues involved in standard setting for assessments like NAEP. Unfortunately, the NAE evaluation of the NAGB standard-setting effort suffers from at least two major weaknesses, and as a result, the conclusions drawn are not well supported by the evidence that is provided. First, the NAE report generally accepts the results and conclusions of the studies it commissioned at face value, even though these studies contain serious flaws, some of which are mentioned in the NAE report. Second, the NAE report evaluates the NAGB achievement levels in a vacuum, in that it fails to consider alternative explanations for anomalous results and fails to examine the potential problems in the methods that it proposes for reporting NAEP results.

Problems in the NAE Studies

The studies commissioned by the NAE Panel contain a number of methodological problems. In the second section of this report, I provide some comments on the specific studies. In this section, I will discuss two general weaknesses that occur in several of the studies. The final report of the NAE evaluation acknowledges some problems in some of the studies, but generally accepts the results and conclusions of most of the studies.

The studies involved in the NAE evaluation subjected the process and products of the NAGB achievement-level-setting (ALS) process to a number of serious challenges, and based on the finding that the ALS results did not meet all of the expectations embodied in these challenges, the NAE Report concludes that the process was fundamentally flawed and that the results are unreasonable. Methodologically, this general approach is consistent with our current views of validation, but there are several serious problems with NAE's implementation of this approach.

Many of the challenges to the ALS process are based on inaccurate assumptions about the purposes of the methodology used in the ALS process. The

Angoff procedure is not designed to achieve consensus in the sense assumed in most of the NAE studies: it is a method for systematically collecting judgements on performance standards and combining these judgements into an overall cutpoint on the scale, and it is neither expected nor required that the judges reach consensus. In using the Angoff procedure, it is not necessary to assume that judges can or will make the transition from a single point on a latent achievement scale to accurate estimates of the associated expected p values, but several of the criticisms of the ALS in the NAE studies and in the NAE report seem to assume that the Angoff procedure depends on this assumption. The NAE Report acknowledges, on page 43, that Angoff judges are not expected to estimate p values for borderline candidates, but yet places heavy reliance on analyses that make this assumption about the Angoff procedure.

A second weakness in the NAE studies is that many of these studies employ methods that are less well researched, and as a result less well understood, than the Angoff procedure as the basis for empirical checks on the Angoff results. In one case, the NAE Panel decided after the fact that a procedure used in their studies (i.e., the item mapping technique) was seriously flawed, but in most cases, the alternative methods are accepted without any built-in checks on their validity. An important example of this general failure to subject the alternative procedures to serious scrutiny is the faith placed in the result of the contrasting-groups method. This method has been around for a long time, but it has not been used all that much, and therefore its properties have not been as thoroughly examined as those of the Angoff procedure. As indicated later in this report, I think that NAE's implementation of this method as a check on the NAGB results was fatally flawed.

An argument based on weak or flawed studies might still be considered convincing if all of the results, or at least the great preponderance of the results, pointed in the same direction. But the NAE studies are not particularly consistent, and a number of the stronger NAE studies tend to support the legitimacy of the ALS process and the reasonableness of the resulting NAGB cutpoints. For example, the international comparisons reported by Beaton and Gonzalez certainly do not suggest that the NAGB achievement levels were set too high, and McLaughlin's comparison of the NAGB cutpoints to cutpoints set by a different panel using a holistic assessment of completed mathematics booklets supported the reasonableness of the link between the final version of the NAGB achievement-level descriptions and the NAGB cutpoints in mathematics.

Implications of Anomalous Results

The second major weakness in the NAE report is that it evaluates the NAGB effort to define achievement levels in a vacuum. There is no serious evaluation of the methods, in particular the anchor-point methodology, that would be used to report NAEP results if the achievement levels were not used. As noted in some of my

comments on the individual NAE studies, many of the criticisms of the NAGB methodology would apply as strongly, or more strongly, to the reporting of results in terms of anchor points. The NAE report makes repeated note of the inconsistencies between current and evolving conceptions of content standards, the NAEP Frameworks, and the NAEP item pools. There are indications, in the contrasting-groups study, of a major inconsistency between NAEP performances and performance as rated by researchers and teachers, possibly indicating a tendency for NAEP to underestimate achievement. These problems were not created by the ALS process and will not go away if anchor points are used, instead of the achievement levels, as the main basis for reporting NAEP results.

The anchor-point method encourages the consumers of NAEP results to make generalizations from individual items to some domain of similar items, without providing any indication of what domain is appropriate, or any evidence to support the legitimacy of such generalizations. What inference should be drawn from the fact that most students at some score level can answer a specific word problem about the area of a rectangular room? Should we infer that students at this score level can generally answer questions about area, questions about rectangles, any question about geometry, or any word problem? The generalizations to be made from anchor items are highly ambiguous. The ALS process has provided the achievement levels as an indication of the kinds of generalizations to be made from the results based on the cutpoint and some evidence to support such generalizations.

A second way in which the NAE studies ignore the context in which the ALS process was implemented is in attributing almost all anomalous results to problems with the ALS process, even when there are obvious alternative explanations. For example, the differences in results for extended-response items and short-answer, dichotomously-scored items are immediately taken as showing inconsistency on the part of the panelists in the ALS process. However, given that the extended-response items are intended to tap different aspects, or dimensions, of achievement, and given that the IRT methodology used to scale NAEP results is based on a strong unidimensionality assumption, it would seem reasonable to at least consider the possibility that these differences might be due to artifacts or limitations in the scaling process. Such alternative explanations are not even mentioned in most of the NAE studies and in the final report.

Concluding Remarks

I think that the evidence provided in the NAE report and in the studies commissioned by the NAE Panel do not provide adequate support for the strong conclusions in the report. The conclusion that the Angoff procedure presents judges with an unmanageable task is based on unwarranted and unreasonable assumptions about what the Angoff procedure is designed to do. The NAE studies attack a straw man when they claim that the Angoff procedure is, "fundamentally flawed," because

the ALS panelists exhibited variability in their ratings for different items and because the panelists did not achieve consensus over the three rounds of the rating process.

The conclusion that the standards that resulted from the ALS process are unreasonably high is based mainly on the results of the contrasting-groups study, which has, I think, serious problems. Taken as a whole, the collection of studies examining the reasonableness of the cutpoints give conflicting results. The international comparisons suggest that the cutpoints may be too low. The comparison with the Kentucky system suggests that the cutpoints are about right. The comparison with AP results suggest that the advanced level at 12th grade may be too high. The comparison with the SAT is ambiguous, because we do not have any clear criteria for what should be considered an advanced performance on the SAT. As a group, these studies are a mixed bag and certainly do not support a strong conclusion that the standards are too high.

Section 2 - Comments on Individual NAE Studies

This section discusses the individual studies, one at a time. Since I have not had an opportunity to review the original data for these studies, my comments are based mainly on the content of the reports of the individual studies.

An Evaluation of the 1992 NAEP Reading Levels—Report One: A Commentary on the Process. Pearson, D. and DeStefano, L. June 1993. (1993a)

This paper provides an extensive qualitative analysis of the process used in developing the achievement-level descriptions for reading. On the whole, this study seems credible. It makes reasonable assumptions about the Angoff procedure. The representation of the Angoff procedure in this study is much more reasonable than that in some of the other NAE studies.

It also provides what appears to be sound advice on how to improve the procedures being used, in particular ways to get better input from public members. The suggestion to have drafts of the descriptors developed first, with input from experts, and then to use the Angoff to refine the descriptors a bit and to set the cutpoints, seems like a good way to avoid potential problems when revisions of the descriptions are made after the cutpoints have been set.

The concern expressed in this paper to the effect that the descriptors and the exemplar items are not fully consistent with the framework is a legitimate criticism. However, it seems evident that this criticism is not a problem with achievement levels, but is a more fundamental issue.

An Evaluation of the 1992 NAEP Reading Achievement Levels—Report Two: An Analysis of the Achievement-Level Descriptors. Pearson, D. & DeStefano, L., June 1993. (1993b)

This report employs numerous sources of data, including observations at various meetings, surveys of meeting participants and input from an expert panel to examine three issues: (1) the quality of the St. Louis descriptors as a basis for level setting in reading, (2) the differences between the St. Louis descriptors used to set cutpoints and the final descriptors used to report on the achievement levels, and (3) the quality of the final version of the descriptors as a basis for reporting the results of the 1992 NAEP in reading (Pearson & DeStefano, 1993b, p. 3). The analysis is in the qualitative tradition and appears to be well done, given its basic assumptions.

The authors take agreement with the Reading framework as the main, if not the only, criterion for evaluating the St. Louis descriptors. It seems to me that this is an overly narrow view of the task given to the panelists in St. Louis. If NAGB were primarily interested in getting achievement-level descriptors and cutpoints that correspond to the Reading Framework, it would have been much more sensible to employ a panel of experts with a detailed knowledge of the Framework, rather than a broadly representative panel of teachers, other educators, and members of the public. One certainly wants the descriptors to be consistent with the Framework (i.e., the content standard) and with the general content of the reading assessment, but the level of achievement (i.e., the student performance standard) specified in each descriptor and its corresponding cutpoint should also reflect the values and opinions of the panelists. That is why they were selected to be representative of teachers and the public. So, I would not see it as a great problem if panelists relied to some extent on a "combination of experience, intuition, and opinion" (Pearson & DeStefano, 1993b, p. 7).

A second and more serious charge is also made by Pearson & DeStefano (1993b, p. 8) that: "the St. Louis descriptors were not consistent with the definition of a good reader put forth in the framework. Instead, they reflect a traditional skills based approach to reading that is contradictory to the philosophy and content of the 1992 NAEP in reading." The experts rejected the St. Louis descriptors as contradictory to current thinking about the reading process (Pearson & DeStefano, 1993b, p. 10).

Pearson & DeStefano (p. 11) conclude with a strong rejection of the St. Louis descriptors:

We conclude that extreme inconsistencies (even contradictions)

between the St. Louis descriptors and the Reading Framework render them invalid as a basis for level setting. The St. Louis descriptors reflect an out-dated, skills-based, developmental approach to reading. By differentiating between achievement and grade levels on the basis of skills, the St. Louis descriptors are antithetical to the basic underpinnings of the framework, namely that the reading process is common to all readers and that differences are based on variations in text and task.

However, it appears that much of the concern expressed by the expert panel employed in this study and incorporated in the conclusions is not limited to the St. Louis descriptors but rather to any performance-based or content-based interpretation for NAEP reading:

It was concluded by the expert panel that the very task that St. Louis participants were asked to complete, to describe performance levels in reading using the framework, was antithetical to the basic assumptions about reading that underlie the development of the framework. Taken seriously, their concerns call into question the very idea of describing levels of performance in reading. At the very least they raise questions as to what the dimensions of the descriptors might be—favoring descriptors that reflect variations in text and task over those that differentiate aspects of the reading process by grade or performance level.

Much of the problem here seems to be that the NAE experts think that the NAEP tests reflect a more traditional conception of reading than the new Framework. However, it is hard to judge the accuracy of the conclusions in this kind of analysis without repeating the analysis or one like it.

An Evaluation of the 1992 NAEP Reading Achievement Levels—Report Three: Comparison of Cutpoints for the 1992 NAEP Reading Achievement Levels with Those Set by Alternate Means. Pearson, D., & DeStefano, L., May, 1993 (1993c).

This paper compared the cutpoints set in St. Louis to a new set of cutpoints that were based on the final version of the achievement-level descriptions cutpoints and were developed using an alternative methodology involving item maps. I think that the alternative methodology for standard setting used in this study is fatally flawed. The results obtained using this methodology necessarily depend on an arbitrarily chosen constant, and therefore in spite of extensive input from the panel of experts, the results are essentially arbitrary.

The methodology used in this study employed item maps as a way of providing a "visual representation of the relative difficulty of all items for each grade level arrayed on the NAEP performance scale" (Pearson & DeStefano, 1993c, p. 5). For dichotomously scored items, a photocopy of the item was placed on the "scale score at which 80% of the students who achieved that scale score answered the item correctly" (Pearson & DeStefano, p. 5). (The constructed response items were treated in a way that was inconsistent with this approach, but that is a relatively minor matter.) The problem is that there is no particular reason for choosing 80% rather than 75% or 85%, 70% or 90%. The choice of 80% is arbitrary.

Yet the 80% criterion is going to determine where items fall on the NAEP scale. If we chose 70% as the criterion, all of the items would move down on the scale. If we use 90%, all of the items would slide up on the scale. The judges set the cutpoint using the item maps. The participants "were asked to consider the description, the reading passages, and the items, and to identify the point at which item difficulty and content was indicative of borderline performance at each achievement level" (Pearson & DeStefano, 1993c, p. 7). The general pattern of where items fall in relation to each other is likely to remain relatively fixed for different values of the percentage-correct criterion (70%-80%-90%), but where the various groups of items fall in relation to the NAEP scale will depend on the value of the percentage-correct criterion chosen.

So, the choice of the cutoff score resulting from the item-mapping technique is likely to be quite sensitive to the choice of the percentage-correct criterion, and this criterion is essentially arbitrary. Therefore, the results of this standard setting process are essentially arbitrary.

Most of the anecdotal information included in Pearson & DeStefano(1993c) is also subject to the same flaw in procedure. For example, the long quote from one of the panelists, which is presented on p. 9, nicely illustrates the problem:

Look here for eighth grade basic. If we look around 244 where the other cutpoint is—most of the items deal with literal comprehension. You have to go up the scale a ways before you get to items that even begin to address the most simple interpretations, inferences, and authors intent. There is nothing that even deals with making predictions.

If they switched from an 80% criterion to a 70% criterion, those items on interpretation, inferences and intent that are "up the scale a ways" would move down the scale. There is likely to be some percentage correct criterion for which these items would cluster around 244.

An interesting footnote here is that the cutpoints set by the alternate procedure

are generally higher than the St. Louis cutpoints. At the advanced level, the alternate method proposed by Pearson & DeStefano (1993c) are 17 points (4th grade), 23 points (8th grade), and 33 points (12th grade) higher than the St. Louis cutpoints.

Rated Achievement Levels of Completed NAEP Mathematics Booklets. McLaughlin, D.H., May 1993. (1993a)

This study focuses on the "effectiveness" of the achievement-level descriptions in mathematics, and addresses two specific questions. First, are the cutpoints reasonable, given the achievement-level descriptions? And, second, would cutpoints based on the revised Nantucket achievement-level descriptions be comparable to the cutpoints based on the original St Louis achievement-level descriptions?

A panel similar to the St Louis panel reviewed 160 math booklets from the 1992 NAEP, with responses from eighth graders. Each panelist sorted 20 booklets into the four categories: below basic, basic, proficient, and advanced based on the achievement-level descriptions. Each panelist therefore implicitly set three cutpoints for the achievement levels using a holistic assessment of completed student booklets. Half of the group used the original St Louis achievement-level descriptions, and half used the revised Nantucket achievement-level descriptions.

The results indicated that the NAGB cutpoints are reasonable, given the achievement-level descriptions, and, that cutpoints based on the revised Nantucket achievement-level descriptions would be comparable to the cutpoints based on the original St Louis achievement-level descriptions. According to McLaughlin(1993, p.10), "...given the very substantial differences in samples and in procedures, these results do not, in themselves, cast a serious shadow of doubt on the official NAEP achievement levels".

On the comparability of the St. Louis and Nantucket achievement levels, McLaughlin(1993, p.11) states that, "At all three levels, the average of the assigned cutpoints based on the St Louis descriptions were higher, although the individual results were not statistically significant...". On the last page of the report, McLaughlin(1993, p.13) points out that, in a study using the Angoff method conducted the next day, the same panelists set a lower standard for the St. Louis descriptions than their colleagues using the Nantucket descriptions, and concludes that, "This reversal is not surprising if we accept the perspective that the differences are really of a size that would be expected by chance; that is, not statistically significant."

This study constitutes quite a strong confirmation of the NAGB ALS process

for eighth grade mathematics and, by extension, of the process as a whole. A completely independent sample of panelists, using a very different method for setting the cutpoints corresponding to the achievement-level descriptions, with a different group running the study came up with essentially the same result as the original ALS process. This was a very stringent test of the reasonableness of the cutpoints given the achievement-level descriptions. The fact that the NAGB ALS process passed the test so convincingly suggests that the cutpoints are aligned with the achievement-level descriptions in a reasonable way.

Order of Angoff Ratings in Setting Multiple Simultaneous Standards. McLaughlin, D.H., May 1993. (1993b)

This study involved ratings by 24 individuals of items in the eighth-grade mathematics assessment. This study was completed on the day immediately after the study of completed NAEP mathematics booklets and involved the same group of panelists.

The major finding of this study is that having panelists do the ratings for the three achievement levels at the same time did not have a major impact on the outcomes. The differences between the levels were a bit more consistent when the three ratings were done at the same time for each item than they were when all items were rated for one level and then all items were rated for the second and third levels, respectively. McLaughlin (p. 128) concludes that, "...it does not seem warranted...to recommend against the use of the three-ratings-per-item methodology."

A stronger conclusion could be drawn from the data. The fact that the Angoff yielded results consistent with reasonable expectations for the different levels, independent of the order in which the ratings were made tends to suggest that the results were reasonable, given the achievement-level descriptors. That is, the ALS process survived a serious check on its consistency with a comparable standard-setting effort.

The second finding of this study was that the results obtained using the original, St. Louis achievement-level descriptions, were not substantially different from the results obtained using the revised, Nantucket achievement-level descriptions. This is an important result in that it supports the use of the Nantucket descriptions with the cutpoints based on the analyses completed in St. Louis.

The third finding is that this replication of the 1992 ALS yielded results that were very similar to those obtained in St. Louis. The fact that a new group of panelists, using a somewhat different methodology, and under different leadership, came up with roughly the same cutpoints provides very strong support for the

connection between the achievement-level descriptions and the corresponding cutpoints.

The one departure from close agreement with the original St. Louis results is also informative. According to McLaughlin (1993, p. 126):

The comparison between Palo Alto and St. Louis results displays remarkable similarity at the proficient and advanced levels, given the major differences in panelist sampling and training as well as differences in item sampling. The substantially lower basic cutpoint set in Palo Alto is probably due to a specific difference in training. Unlike the St. Louis panelists, Palo Alto panelists had, on the preceding day, viewed actual eighth grade students' booklets during "training" and had seen actual performance dropoffs, including omitted items, toward the ends of item blocks for poorly performing students. Therefore, they tended to estimate lower percents correct for the borderline basic students at the end of the test. Not having expected this, the author did not instruct panelists concerning the treatment of not reached items. In fact, the "not reached" items at the ends of blocks had not been treated by ETS as "wrong" in developing the NAEP scale. As a result, Palo Alto panelists ratings for borderline basic students were lower than those of the St. Louis panelists. Overall, the conclusion to be reached from this comparison is that the two studies yielded similar results.

This is one of several places in the NAE studies where we have evidence of patterns of performance that are somewhat inconsistent with common interpretations of NAEP results and that may introduce bias into the scaling process. The performance dropoffs at the ends of blocks for poorly performing students could easily be due to lack of motivation or lack of time. Neither of these possibilities is taken into account in the usual interpretations, and either of these possibilities would distort the estimation of item parameters in IRT.

Teachers' and Researchers' Ratings of Student Performance and NAEP Mathematics and Reading Achievement Levels. McLaughlin et.al., June, 1993

This study compared the NAGB cutpoints in reading and mathematics to new cutpoints for each achievement level developed using a version of the contrasting groups method for standard setting. The stated purpose of this NAE study was to determine whether the audiences for the NAEP reports will interpret the words, "proficient", "basic", and "advanced", as used in the NAEP reports, "in the same manner that the panelists who set the cutpoints did" (McLaughlin et.al., 1993, pp. 1-2).

The main conclusion drawn in the report of this study seems to be that there was, "...a noticeable discrepancy between these teachers' interpretations of the achievement level descriptions ... and the interpretations by the panel selected by NAGB to set the achievement level cutpoints (McLaughlin et.al., 1993, p. 13). The results reported in the paper do not support this conclusion, because the methodology employed in the study suffers from at least three major problems, each of which is a serious potential source of bias.

First, and most serious, the performances rated as part of the NAE contrasting-groups study were not NAEP performances. In particular, there were two characteristics of the individual assessments administered to students by the researchers in this study that would tend to bias the ratings of student performance upward compared to performance on NAEP, and therefore would tend to lower the cutpoints. First, since these assessments were administered one-on-one, it is likely that all of the tested students would make a serious effort to answer the questions as well as they could. This would be likely to lead to better performance than that observed on NAEP, because there is some evidence that at least some students do not try as hard as they might on NAEP (e.g., McLaughlin et.al., 1993, p. 3, mention that there was some loss of data in this study because some students failed to attempt any items on the NAEP test).

In addition, during the individual assessments by the researchers, multiple prompts or probes were used to elicit responses from students. So, for example, the instructions for the reading assessment suggest that, in questioning students about the passages they have read, one probe "...will usually suffice. But, if you get no response to one, you may want to try one of the others." (McLaughlin et.al., 1993, p. C-37) On the NAEP tests, the students are on their own; if some students misunderstand a question, they do not get a different prompt and then a second or third chance to answer the question. Therefore, the use of multiple prompts/probes in the context of a one-on-one, individualized assessment is likely to produce better performance for most students than would be observed on NAEP. The impact of these two sources of bias are likely to be most pronounced for relatively weak students, but would tend to improve performance to some extent at all levels of performance.

Therefore at least some of the differences between the researchers' ratings of performance and the NAEP results, as reported by McLaughlin et.al., and quite possibly all of these differences, can be attributed to differences in the performances being rated. That is, contrary to the conclusions drawn in McLaughlin et.al.(1993, p. 13), the panels selected by NAGB to set the achievement level cutpoints and the researchers in the contrasting-groups study may have been using exactly the same interpretation of the achievement-level descriptions, and yet gotten different results simply because they were applying the standards to very different kinds of performances.

We do not know exactly what performances the teachers were thinking of when they categorized students as basic, proficient, and advanced, but we do know that the teachers consistently classified even more students as being above each cutpoint than the researchers. It is likely that the teachers are classifying students in terms of what they are capable of doing, i.e., in terms of their best performances over the school year rather than in terms of their performance on specific assessments like NAEP.

This first source of bias certainly has some impact on the results, and may very well explain all of the differences reported in McLaughlin et.al. (1993). Now, one might argue that even if the results are due entirely to difference in the performances being rated, the fact that the teachers seem to be interpreting the results in ways that are not consistent with the kinds of performances included in the NAEP suggests that these teachers are making inappropriate interpretations of the results. There is clearly some logic in this point of view, but note that this argument casts doubt on any content-based interpretation of NAEP results, and does not indicate any specific problem in the achievement-level descriptions.

The second major methodological problem in the contrasting-groups study reported by McLaughlin et.al. (1993) is the failure to include any consideration of the differences in the reliabilities of the various assessment methods being used. The errors of measurement that are associated with unreliability cause observed score distributions to have a greater variance, and therefore more extreme scores, than the corresponding true score distributions. NAEP results focus on the distributions of true scores in each content area at each grade level. The results reported for the contrasting-groups analysis are based on observed scores, and are therefore likely, for this reason alone, to have a higher frequency of extreme scores, particularly at the advanced level than they would if the classification were error free.

The data in McLaughlin et.al. (1993) indicate that the classifications in the contrasting-groups study were not very reliable. The classifications of students by the teachers and researchers did not agree very closely in reading, with correlations of about 0.70. The correlations in mathematics were much lower, with values around 0.30. These correlations suggest that the reliability of the teachers' ratings, or the reliability of the researchers' ratings, or both, were not very high, especially in mathematics. Therefore, the impact of the failure to consider the reliabilities of the classification procedures may have introduced substantial bias into the results.

The mechanism whereby the unreliability of the classifications made by teachers could introduce bias into the results can be illustrated by considering the boundary between the proficient and advanced categories. Because the classification method is unreliable, some students who should be classified as proficient are classified as advanced. Some advanced students will also be misclassified into the proficient category, but this will happen less often because there are fewer students

who are in the advanced category than there are in the proficient category. As a result, when the distribution of NAEP scores for "advanced" students is examined, it will appear that there are substantial numbers of "advanced" students with relatively low NAEP scores, thus suggesting that the cutpoint for the advanced category should be set lower than it otherwise would be. A similar pattern will tend to move the cutpoint for the basic category a bit higher than it would otherwise be, but this effect may have been swamped by the effect of using assessment procedures (e.g., one-on-one assessment, with multiple probes) that tend to enhance performance, especially the performance of weaker students.

It could be argued that the whole idea of using analyses like those reported in Tables 1 - 4 of McLaughlin et.al. (1993) is highly questionable. It has been consistently maintained that NAEP results for individuals are too unreliable to be used for most purposes. Not only are the contrasting-groups analyses using the NAEP scores of individuals in a way that depends on their being fairly reliable, but by associating five random normal imputations with each score, McLaughlin et.al. have made the individual NAEP scores even less reliable than they were to begin with (note that the imputation method used by McLaughlin et.al. is different from that used by ETS to generate NAEP distributions).

The third potential source of bias in the results of the contrasting-groups study is in the methods used to equate scores on the 1993 NAEP pilot tests to scores on the 1992 NAEP assessment. There are not enough details in the report to be completely clear about what was done, but the equating of the 1993 booklets to the 1992 booklets appears to rest on rather shaky foundations. There were 489 scores available on the 1992 booklets included for the study, distributed over two grade levels and two content areas, thus leaving about 125 scores on 1992 booklets for each grade level and content area (McLaughlin et.al., 1993, p 3). It seems that two 1992 forms were used in each case, further reducing the average sample size for each equating to about 65. This sample size does not seem large enough to get a stable equating relationship using equipercentile equating. The instability introduced into the equating by the small sample sizes is likely to be most severe for extreme scores, where the population frequencies are low to begin with. The strange patterns in the correlations among the 1992 booklets and the 1993 booklets, reported in Appendix B (McLaughlin et.al., 1993, p. B8) may be due in part to the small sample sizes being used. The weaknesses in the equating do not necessarily introduce bias in any particular direction, but will tend to make the results unstable.

A particularly disturbing aspect of the contrasting-groups study is the failure to examine, or at least to report any analyses, of some highly anomalous results. There are a number of tables in Appendix B that report crosstabs between classifications of students by teachers and classifications based on the student's imputed NAEP scores. In these tables, there are substantial numbers of entries for which the teacher's classification is at the advanced level but the imputed score falls below the cutpoint

for basic achievement. Such large discrepancies are worthy of some discussion. Did the students fail to respond to some parts of the NAEP booklet, perhaps because of a lack of time or motivation? Did the normally distributed imputation process have an undue influence on some scores? Were the teachers overgenerous? Large numbers of outliers can indicate serious problems in a data base or in the choice of analysis, and therefore deserve careful examination. There is no indication in McLaughlin, et. al.(1993) that any of these potentially serious problems were examined.

Validity of the 1992 NAEP Achievement-Level-Setting Process. McLaughlin, D.H., May, 1993 (1993c)

This paper reports on a number of analyses designed to examine the validity of the ALS process by looking for certain patterns in the data. In particular McLaughlin divided the results from the ALS process in various ways (e.g., hard items vs. easy items, MC vs. short answer) and assumed that any differences found between different parts of the data base constitute evidence against the validity of the ALS results. Most of these analyses did not yield the kinds of differences that McLaughlin was looking for, but several substantial differences were found.

McLaughlin's paper contains several analyses that raise serious questions about appropriate interpretations for NAEP results. However, the general form of the argument in the paper and the conclusions that are reached based on the data presented are not appropriate. There are two major flaws in this paper. First, McLaughlin claims that the Angoff procedure makes certain assumptions that it does not make. Second, McLaughlin ignores a host of possible explanations for the results of certain analyses, in favor of explanations that are consistent with the general thrust of his argument. I will discuss these two points in turn and then give some attention to several unresolved issues that deserve further attention.

Assumptions Implicit in the Use of the Angoff Procedure

The Angoff procedure is basically a systematic procedure for collecting judgements about where standards should be set on examinations. Judges are told to think about a marginally qualified examinee and to record the probability that such an examinee would get each item correct. The point of this exercise is to help the judges to make their general sense of what would constitute appropriate standards in a given context more explicit. It is not to train the judges to estimate p values or IRT parameters for any group of students. The purpose of all of the standard-setting methods currently in use (including the Angoff and the contrasting groups method) is to help the judges to set reasonable standards by making a highly abstract issue (i.e., decisions about how much is enough) a bit more concrete, and perhaps a bit more manageable.

McLaughlin(1993c, p. 80) states that "The major assumption of the Angoff process is that panelists can imagine students in the reference population who are at a specific performance level (e.g., the borderline between basic and proficient performance) and can, accurately or approximately, estimate the proportion of those students who would respond correctly to each item on the assessment." This view may seem somewhat plausible because of the way that the Angoff procedure is usually described, but it is fundamentally wrong. The Angoff procedure is not intended as a way of estimating p values, and therefore should not be evaluated primarily in terms of how well it estimates p values. If we want estimates of p values or IRT parameters, we can get those directly from student performance data.

But McLaughlin (1993c, p. 78) imposes an even stricter set of requirements on the Angoff procedure. He claims that if Angoff ratings of different types of items yield different cutpoints on the NAEP scale, "then either panelists are not imagining a single point on the NAEP scale for each level (when looking at different items), or they are not able to translate accurately from an underlying scale point to the percent of students who would get a particular item right". This is not what panelists were asked to do and it is not what they are expected to do. The panelists were chosen to provide broad-based input to a policy decision: they were not expected to implement IRT models in their heads.

On page 101, McLaughlin states his expectations even more dramatically:

A fundamental question must be addressed: Can panelists be found and prepared who can validly conceive of NAEP achievement levels, and, moreover, validly estimate the percent of students at an achievement level who would get a test item right? If not, then the Angoff procedure must be abandoned for this task.

It seems that McLaughlin expects the panelists to mimic a fairly complex set of computer routines.

McLaughlin(1993c, p. 79) also claims that: "The acceptability of the achievement levels depends on the demonstration that there is consensus among the panelists, who were selected to represent the nation". The requirement imposed by McLaughlin (1993c, pp.78-79) that the Angoff procedure lead to consensus among the panelists is unrealistic and inappropriate. A broadly representative group of panelists were used so that the achievement level descriptions and corresponding cutpoints would reflect the views of different groups. Differences of opinion about the standards to be adopted were expected. The use of multiple rounds in the ALS process was designed to give the panelists adequate opportunities to reflect on the choices they were making. The feedback provided to panelists was designed mainly to give panelists a chance to reconsider judgements that seemed to be outliers. The process was not designed to reach consensus.

Interpretations of Results

There are many possible interpretations that could be assigned to the results of those studies reported by McLaughlin that yielded significant differences. McLaughlin(1993c) basically dismisses any possible interpretation other than those that relate to his main thesis - that the ALS process did not work. Now, the ALS process is new and has encountered several unanticipated problems. These problems should be recognized as such, and the proposed solutions should be evaluated in an evenhanded way. To automatically assume that every anomalous result is due to faults in the ALS process is unfair, and may cause us to overlook other possible sources of misinterpretation, and thereby to miss opportunities to improve the overall NAEP program.

In discussing the differences in cutpoints obtained for extended-response and dichotomously-scored items, McLaughlin takes it as a given that these differences are due to methodological flaws in the ALS process or to "inherently different achievement levels". In doing so, he ignores other possible explanations for the differences that are, I think, at least as plausible. One alternative explanation is that performance is substantially different for these two item types. For example, it is certainly possible that a lack of motivation could effect the two item types differently, especially for low-scoring students, because it takes more effort to respond to an extended-response item. To the extent that this phenomenon is occurring, it could explain some of the differences between the two item types.

A second alternative explanation would be an artifact in the scaling process. A major argument for including extended-response items is that they measure something beyond MC items. If this is so, that is, if the two item types tap different kinds or dimensions of achievement, then the basic assumption of IRT theory is violated, and it is at least possible that artifacts in the scaling process could explain the differences observed. (In a sense, the possible problem with motivation is a special case of this more general problem, with the second dimension being motivation.)

Similarly, in discussing differences in cutpoints resulting from different kinds of short-answer items, McLaughlin(p.87) states that:

If such differences are found, they may be due either to panelists' inability to translate achievement levels into quantitative item performance predictions, which is, after all, an extremely difficult task, or they may be due to the panelists' imagining different ability factors. In either case, the assumption of the judgement is not met.

The anomalous finding reported in this section of the McLaughlin report is the difference between hard items and easy items. (The reported difference between MC

and short answer items is confounded with this hard-easy difference). McLaughlin suggests that this difference may be due to a failure of panelists to take difficulty into account. No other possible explanation is considered. Yet there are many other plausible possibilities. The pattern in the data immediately suggests a regression artifact, and the IRT models used in scaling NAEP are basically nonlinear regressions of observed performance on an unobservable latent ability. I am confident that if one used only "easy" items to estimate scores on the NAEP scale, one would not get the same result that one would get if one used only "hard" items.

McLaughlin's statement on page 91, that:

There is no plausible alternative to the conclusion that this task required skills in item interpretation that were beyond the reach of the participating panelists, in the context of the training and instruction they were given.

is simply wrong. There are plausible alternative explanations for the anomalous findings reported by McLaughlin, and the Angoff task does not require all of the skills that McLaughlin says it requires.

Unresolved Issues

There are a number of issues in the McLaughlin report that clearly deserve further attention. Most of these issues have been discussed by the ACT Technical Advisory Committee for Standard Setting, which made some suggestions for how to deal with these issues, but more work is needed.

A major problem is how best to set standards for extended-response items. The inconsistency between short-answer items and extended-response items is a practical problem because it introduces a kind of sampling error into the estimates of cutpoints. Probably more important, it raises conceptual questions that need to be answered, in particular, questions about what is causing the discrepancy. We have little experience setting standards for extended-response items, and I think that there are a number of fundamental problems in this area.

The difference in results for the hard and easy items also deserves further attention. I think that this difference may be an artifact of the scaling procedures used for NAEP, but it could also be a problem in the ALS process.

Comparisons of Student Performance on NAEP and other Standardized Tests.
Hartka, E., 1993

This paper basically compares proportions in the advanced category on NAEP to estimates of the proportions achieving relatively high scores on the SAT or the AP exams. The results suggest that the cutpoint for the advanced level is pretty high in a norm-referenced sense, but the advanced level was supposed to be high. However, the results for the SAT involve ambiguity in what constitutes advanced performance on the SAT.

Comparing the NAEP Trial State Assessment Results with the IEAP international Results. Beaton A.E., & Gonzalez, E.J., August, 1993

By relating results on NAEP and the IEAP, Beaton and Gonzalez, are able, based on a number of assumptions, to estimate the percentages of students in a number of foreign countries who would be at each NAGB achievement level if they took the NAEP. According to Beaton and Gonzalez (1993, p.20) "Perhaps the most striking observation from this table is the large percentage of students in Taiwan(24.4%), and Korea(15.6%) who have reached the advanced level". The data in this report suggest that the NAGB cutpoints are not particularly high when compared to performance in those countries with the best performance on the IEAP.



National Assessment Governing Board

National Assessment of Educational Progress

FOR IMMEDIATE RELEASE
December 2, 1997

Contact: Lawrence Feinberg
(202) 357-6942

National Assessment Governing Board Plans to Develop Voluntary National Tests Based on NAEP Assessments

The National Assessment Governing Board has set out plans for developing voluntary national tests in 4th grade reading and 8th grade mathematics, based on the National Assessment of Educational Progress (NAEP).

Under the Appropriations Act for 1998, passed by Congress and signed by the President in mid-November, the Board has been granted "exclusive authority" over development of the proposed exams. The law prohibits any pilot testing of questions or field tests before October 1, 1998, and also provides for studies and evaluations by the National Academy of Sciences.

At its regular quarterly meeting in Washington on November 22, the Board adopted procedures for making decisions about the contract and specifications for developing the test questions and design. It promised wide public input into its deliberations.

"We will carry out the job Congress has asked us to do--develop an individualized version of the National Assessment of Educational Progress," said Mark D. Musick, chairman of the independent, bipartisan 26-member Governing Board, which has set policy for NAEP since 1988. The proposed tests would show students whether they can meet NAEP's standards for basic, proficient, or advanced performance.

According to a House-Senate conference report, the new tests, designed for individual students, are required to "be based on the same content and performance standards as are used for NAEP," which is given to groups of students to measure academic achievement. Congress directed that the tests be "linked to NAEP to the maximum extent possible," which the Board will try to make sure is done.

Over the past 27 years, results from National Assessment exams have been reported for the nation, regions, and states, but not for individual students.

"The Board plans to act deliberately and fairly," Musick said. "We will hold public hearings. We will listen carefully. We want to give plenty of opportunity for a wide range of views to be considered."

-MORE-

800 North Capitol Street, N.W.
Suite 825, Mailstop 7583
Washington, D.C. 20002-4233
Phone: (202) 357-6938
Fax: (202) 357-6945

Musick added: "We are aware of the difficulties and challenges that are part of the directives from the President and Congress. We are not unaware of the complexity and controversy surrounding the voluntary national tests. We believe that this task was given to the Board because it has proven since 1988 that it is fair, bipartisan, deliberative, and independent as it tackles complex and controversial issues."

The Board's plans include the following:

- A special committee will review the five-year, \$65 million test development contract awarded by the Department of Education on August 15 to a group of major test publishers, headed by American Institutes for Research. The committee, composed of seven Board members, will make recommendations to the full Board. Under the law, NAGB has until February 11 to approve, modify, or terminate the contract. If the contract is rescinded, the Board has authority to negotiate a new one.

- A special meeting of the Board will be held on January 22 in Washington to deal with national test development issues.

- The Board's two academic committees will begin the process of developing specifications for the reading and mathematics tests by reviewing recommendations of the expert panels convened last summer by the Department of Education. The panels were set up under a contract with the Council of Chief State School Officers and MPR Associates. On September 25, Education Secretary Richard Riley announced that he would not approve the specifications proposed by the panels or forward them to the test development contractor.

- The Board will hold public hearings on test specifications this winter. It aims to adopt final specifications at its meeting scheduled for March 7 and 8 in Seattle, WA.

- The Board will review all test questions to make sure they cover the same content and meet the same standards of appropriateness and being free from bias that are used in NAEP.

- The Board will meet with representatives of the National Academy of Sciences, which will be conducting three Congressionally mandated studies related to development of the national test. The Academy studies deal with (1) the feasibility of developing a common scale to compare commercial and state tests and NAEP; (2) evaluation of national test items and design for technical quality and validity and whether they "can be used for tracking, graduation, or promotion of students;" and (3) recommendations to avoid discriminatory use of existing and new tests. The reports are due to Congress by September 1, 1998, with an interim report on the equating study due June 15.

- The Board will consider policies for test reporting and use through a process that ensures wide consultation and input.

-MORE-

- The Board will seek to complete all preliminary development work by September 30, 1998, so there can be pilot testing and field testing later, as determined by law.

"The Governing Board has not endorsed a voluntary national test and we have not opposed a voluntary national test," Musick said. "The Board is doing what Congress has asked us to do.

"The position of the Board has been that if there is to be a test that essentially is an adaptation of NAEP, then it should come under the NAEP umbrella," he continued. "If the country finds that useful and Congress funds it and moves forward, we hope to develop a test that is worthy of consideration."

The contract review committee, which begins work immediately, is headed by William T. Randall, a former NAGB chairman who was state Commissioner of Education in Colorado for many years. Other Board members on the committee include vice-chairman Mary Blanton, of Salisbury, NC, a lawyer and general public member of NAGB; James Ellingson, of Moorhead, MN, a fourth grade teacher; Thomas Fisher, director of student assessment for the Florida Department of Education; Edward Haertel, a professor at Stanford University, who is an expert in educational testing; and Deborah Voltz, assistant professor of special education at the University of Louisville.

Also serving is a new Board member, Diane Ravitch, former Assistant Secretary of Education under President Bush. Ravitch is a senior research scholar at New York University and a senior fellow at the Brookings Institution.

The National Assessment Governing Board, established by Congress in 1988, includes governors and state legislators of both parties, local school officials, teachers, testing experts, and business and public representatives. The Board sets policy for NAEP, including test content and performance standards.

Since 1970, the National Assessment, based on nationally-representative samples of students, has been the only source of reliable, continuing information about what American students know and can do in various academic subjects. Since 1990, it has provided the only comparable state-by-state data on academic achievement, based on representative samples of students in each participating state. Over the past seven years, 48 states have participated in NAEP and agreed to have results reported according to NAEP performance standards.

###



National Assessment Governing Board

National Assessment of Educational Progress

FOR RELEASE
Friday, November 21, 1997

Contact: Lawrence Feinberg
202-357-6942 or
(11/21-22) 703-412-0053

NATIONAL ASSESSMENT GOVERNING BOARD PLANS TO DEVELOP NATIONAL TESTS BASED ON NAEP ASSESSMENTS

Statement by Mark D. Musick

Chairman, National Assessment Governing Board;
President, Southern Regional Education Board

The National Assessment Governing Board today outlined plans to handle its new responsibilities to develop the tests for the voluntary national test program. As you know, the Board was given "exclusive authority" over the test development contract in a law passed by Congress and signed by President Clinton last week.

Our aim will be to carry out the intent of Congress, which is clear in the House-Senate conference report, that these new tests, designed for individual students, "will be based on the same content and performance standards as are used for NAEP," which is a survey of how groups of students perform. The Congress also directed that the tests be "linked to NAEP to the maximum extent possible," which the Board will try to make sure is done.

The Board plans to act deliberately and fairly. Our plans include the following:

- A special committee will review the test development contract previously awarded by the Department of Education and make recommendations to the full Board. Under the law, the Board has until February 11 to approve, amend, or terminate the contract.
- A special meeting of the Board will be held in mid-January to deal with national test development issues.
- The Board's two academic committees will review the proposed specifications for the tests of 4th grade reading and 8th grade mathematics. The Board will hold public hearings on the test specifications and aims to act on them at its March meeting;

800 North Capitol Street, N.W.
Suite 825, Mailstop 7583
Washington, D.C. 20002-4233
Phone: (202) 357-6938
Fax: (202) 357-6945

- The Board will review all of the test questions prepared by the contractor to make sure they cover the same content and meet the same standards of being appropriate and free from bias that are used in the National Assessment of Educational Progress.
- The Board will meet with representatives of the National Academy of Sciences, which will be conducting Congressionally mandated evaluations related to development of the national test.
- The Board will seek to complete all of the work directed by Congress by September 30, 1998 so that there can be pilot testing and field testing later.

The National Assessment Governing Board was established by Congress in 1988 as an independent, bipartisan board, composed of state and local officials, educators, lay citizens, and business leaders. Our new responsibilities for the voluntary national tests are not the first challenges that the Board has been given.

Congress directed the Board in 1988 to change an assessment that had surveyed only national samples of students into a program that could give individual states a fair and independent measure of academic achievement. The Board was asked to develop "appropriate" standards of student achievement for each grade and subject tested. Since 1990, 48 states have participated in the state-by-state National Assessment program, which has been carried out in mathematics, reading, and science.

Now President Clinton and Congress have directed the Governing Board to take an assessment that has been used voluntarily by nearly every state to get a state-level measure of achievement and develop an individual version of NAEP that will give information to parents, teachers, and students.

At the Board's regularly quarterly meeting today we have created a special committee to review the test development contract. As the new law requires, the contract will be carefully reviewed, but under the law, the Board has only 90 days--until February 11--to make any contract changes or decide to terminate it. As chairman of this special committee, I have appointed a former chairman of the Board, William Randall, who for many years was state Commissioner of Education in Colorado. The other members include the Board's Vice Chair, Mary Blanton, a general public member of NAGB, and James Ellingson, Thomas Fisher, Edward Haertel and Deborah Voltz. I have also appointed one of the newest members of the board, Diane Ravitch to the committee. She is a distinguished scholar who has written extensively on education standards, and she also was an Assistant Secretary of Education under President Bush.

In mid-January the Governing Board will hold a special meeting to consider the Committee's recommendations and take up other matters pertaining to the national test development.

The Board is also setting in motion a process to adopt specifications for the tests in 4th grade reading and 8th grade mathematics. The process will include a review of the test specifications that have already been proposed. We will hold public hearings. We will listen carefully. And we will carry out the job Congress has asked us to do—develop an individualized version of the National Assessment that will be able to show students whether they can meet NAEP's standards of basic, proficient, or advanced achievement.

We are confident that we can meet the timetable that Congress has laid out. We also will make sure that the Board takes care of its regular work of setting policy for the National Assessment of Educational Progress.

Since the Governing Board was established almost a decade ago, we have worked hard to develop high student performance standards on NAEP, which require the mandate and the commitment that only an independent, broadly-representative Board can give.

We are aware of the difficulties and challenges that are part of the directives from the President and Congress. We are not unaware of the complexity and controversy surrounding the voluntary national tests. We believe that this task was given to the Board because it has proven since 1988 that it is fair, bipartisan, deliberative and independent as it tackles complex and often controversial issues. We believe that we can do what the President and Congress have asked of us.



National Assessment Governing Board

National Assessment of Educational Progress

June 2, 1997

Honorable Richard A. Riley
Secretary
U.S. Department of Education
600 Independence Avenue, S.W.
Washington, DC 20202

Dear Mr. Secretary:

On May 9, Michael Cohen, Assistant to the President for Education, briefed the Governing Board on the President's voluntary national test initiative in fourth grade reading and eighth grade mathematics. During his presentation, Mr. Cohen raised the issue of governance and invited advice from the Board. This letter, which has been reviewed by our Executive Committee, is to serve as an initial response, prompted by his invitation and by inquiries from the Hill and others.

We have carefully considered the proposed initiative. In so doing, we found ourselves returning to one central conclusion: that the voluntary national test initiative is, in essence, NAEP. The proposed tests will be based on NAEP frameworks. Results are to be reported using achievement levels established by the Governing Board. There is an implicit reliance on NAEP's reputation for quality and integrity. The use of NAEP as the basis for the voluntary test initiative is purposeful and makes it difficult to argue that the initiative is not NAEP.

If the voluntary test initiative is NAEP, it follows that it should be included under the NAEP governance and operations structure (comprised of the Governing Board, NCES, and various contractors); as opposed to creating a separate, and possibly competing, structure. We believe that the initiative will have a better chance of success if it is made a part of NAEP and that there are important, related reasons for following this course.

One reason is to protect the integrity of the National Assessment program itself and to enable the voluntary test initiative to benefit from that integrity. The link to NAEP is a fundamental part of the voluntary test initiative, relying on NAEP's challenging content and its standards-based reporting. Some observers worry that this link could have negative consequences for the National Assessment program. For example, participation in the voluntary test initiative could affect the ability to draw representative samples for the National Assessment. This in turn could affect the strength of the comparability of the voluntary national test results to the National Assessment. These two activities are interdependent. Decisions and actions that affect the National Assessment will, ultimately, affect the voluntary national test and vice versa.

800 North Capitol Street, N.W.
Suite 825, Mailstop 7533
Washington, D.C. 20002-4233
Phone: (202) 357-6938
Fax: (202) 357-6945

Coherent policies are needed to ensure over time that results of the National Assessment and the voluntary test initiative are comparable, technically sound, and achieved through procedures that are well coordinated and operationally feasible. We do not see how this could be accomplished in an efficient, cost effective manner, if at all, should policies and operational strategies for the respective activities be developed by separate agencies. Such coherence is most likely where there is a single governance and operations structure and least likely where there are multiple structures.

For example, the Governing Board changes NAEP frameworks periodically. Reading and mathematics frameworks both are nearing the ten-year mark and may need to be revised. Framework revisions could affect test design for NAEP and also for the voluntary reading and mathematics tests, because they are to be based on NAEP frameworks. Obviously, the respective test development and testing cycles, contract schedules, and user needs would all require consideration and balancing. A single governance and operations structure would, as a matter of course, take all of these factors into account as a part of policy development and program planning. A separate structure for the voluntary test initiative would operate under a different set of incentives and assumptions. This could lead to ambiguity of roles, uncertainty about key decisions, and inadequate coordination with NAEP policies and implementation strategies. The added costs and duplication of effort also argue against having a separate structure for the voluntary test initiative.

A temporary "wall" has been erected during the current development phase between the NAEP program and the voluntary test initiative. This has been explained as a means for the Department to move expeditiously on the voluntary test initiative. It is based also on the fact that the current development phase is being supported by separate funding and separate legislative authority and that NAEP lacks specific authority to develop an individual student test. As a result, the Governing Board is not involved, nor has it sought to be involved, in the activities associated with the voluntary test initiative. However, because of our responsibility for NAEP and the inquiries we have received, we are now thinking about these issues. For the sake of the integrity, effectiveness, and coherence of both the National Assessment and the voluntary test initiative, logic and experience argue for the Governing Board to be the policy body for the voluntary national test initiative.

Considerations in whether to have the Board involved in the governance of the voluntary test initiative should be the Board's composition, independence, and experience. The Board is broadly representative of state and local interests. Its members include two governors and two state legislators of both parties, two chief state school officers, a member of a state board of education, a local school board member, a representative of private schools, a district superintendent, elementary and secondary principals, and three teachers. In addition, test specialists, curriculum specialists, a business representative, and members of the general public round out the Board's membership. By law, the Board is independent of the Secretary and the Department of Education.

The Secretary - Page 3

The Board's composition makes it well-suited to develop policies that directly affect state, local and private schools. Its independence could help ameliorate the debate about the national versus federal character of the voluntary test initiative. Its nine-year experience governing NAEP suggests that it has the immediate capacity to set policy for the proposed voluntary national test initiative. A new, separate body would need a significant amount of time for initial appointments, organizing itself, and becoming well-grounded on many varied technical and policy matters.

We appreciate the opportunity to provide these initial thoughts to you. Many of these matters will need to be addressed through NAEP reauthorization. Toward that end, we look forward to continued conversations with you on these and related matters.

Sincerely,

A handwritten signature in cursive script, appearing to read "Bill", is written over the printed name.

William T. Randall
Chairman

cc: Michael Cohen
Marshall Smith

Review Procedure: Test Development Contract for the Voluntary National Tests

PURPOSE

The purpose of this paper is to describe the procedure for reviewing and modifying the Test Development Contract for the Voluntary National Tests.

BACKGROUND

Under the FY 1998 Labor, HHS, and Education Appropriation Act (the Act), enacted on November 13, 1997, the National Assessment Governing Board is granted "exclusive authority over all policies, direction, and guidelines for developing voluntary national tests pursuant to the [contract with AIR]..." The Act provides that no FY 1998 appropriation may be used to pilot test, field test, implement, administer, or distribute in any way, any national tests.

The Act requires, within 90 days of enactment, the Board to review the national test development contract and modify the contract as the Board determines necessary. The Act further provides that "...if the contract cannot be modified to the extent deemed necessary by the Board [and in a manner not inconsistent with the Act], the contract shall be terminated and the Board shall negotiate a new contract, under the Board's exclusive control, for the tests, not inconsistent with this Act or applicable laws."

The conference report accompanying the Act indicates the conferees' expectation that the national tests "will be based on the same content and performance standards as are used for NAEP and...linked to NAEP to the maximum extent possible." Thus, review of the test specifications, because of their bearing on test development, must be a consideration in the contract review process and timed and coordinated accordingly.

ASSIGNMENT

The review of the voluntary national test development will be conducted by a special committee appointed by the Chairman and reporting to the Executive Committee of the National Assessment Governing Board. The special review committee will be assisted by Governing Board staff and appropriate consultants as needed.

NOTE: A special one-day meeting of the Governing Board shall be conducted on January 22, 1998. At this special meeting, the contract review committee shall present its recommendations on the contract for review and action.

PROCESS

Step

Timeline for Completion

1. **Obtain relevant documents**, including the original RFP, copies of the program and cost proposals, copies of responses to negotiation questions, and final award documents. Obtain clarification on NAGB's scope of authority to conduct any needed negotiations directly with the contractor.

Within 7 days of enactment
(Completed November 17)

2. **Provide orientation** to special review committee on the authorizing legislation and the criteria and procedures to be used in reviewing the contract (see 3 below). Discuss the criteria that would result in a termination.

Within 14 days of enactment
(By November 27)

3. **Conduct initial review**

(a) **Review RFP for conformance with authorizing legislation and applicable NAGB policies**. Identify areas of concern and/or non-conformance with the legislation and NAGB policies.

Within 33 days of enactment
(By December 16)

(b) **Identify and list all deliverables** required under the RFP and the due dates, review criteria and timelines for review and/or acceptance of the deliverables. Identify issues and concerns with respect to the deliverables.

Within 33 days of enactment
(By December 16)

(c) **Identify and list requirements under the RFP for advice and/or advisory committees**. Identify the criteria and process for selection and appointment. Identify issues and concerns with respect to advice and advisory committees.

Within 33 days of enactment
(By December 16)

(d) **Review program proposals for conformance with authorizing legislation, RFP, and applicable NAGB policies**. Identify areas of concern and/or non-conformance with the authorizing legislation, the RFP, and applicable NAGB policies.

Within 33 days of enactment
(By December 16)

(e) **Review the cost proposal** taking into account any issues or concerns that have arisen as a result of the reviews in (a) through (d) in this section.

Within 33 days of enactment
(By December 16)

4. **Initial Recommendation**

Determine whether the contract requires modification and whether the required modifications reasonably

Within 57 days of enactment
(By January 9)

can be made within the current contract OR whether the contract shall be terminated.

4. (a). IF the determination is to recommend at this point enactment to terminate, **prepare a draft termination letter.**

Within 64 days of
(By January 16)

4. (b) If the contract requires modifications, **prepare a draft modification letter** listing the issues, concerns, and actions required of the contractor.

Within 64 days of enactment
(By January 16)

4. (c) **Present the draft termination letter or the draft modification letter** to the Governing Board for discussion and action. The Board will give prospective authority to enact- the special review committee and the Executive Committee jointly to act on behalf of the Board in conducting termination proceedings or contract modification negotiations, as appropriate. The Board will identify issues, if any, that must be brought back to the Governing Board for decision.

January 1998 Board meeting (Jan. 22)
(Within 70 days of
ment)

5. **Send final version of termination letter or modification letter to the contractor.**

If termination, go to step 11 below.

Within 71 days of enactment
(By January 23)

6. **Receive contractor's response**

Within 78 days of enactment
(By January 30)

7. **Evaluate contractor's response**, including contractor's evaluation of any cost implications of needed changes, and determine whether modifications of the contract can be made OR whether the contract shall be terminated. If, on the basis of the contractor's response, the special committee and the executive committee determine that the contract cannot be modified to the extent determined necessary by the Board, the contract shall be terminated. Go to step 11 below.

Within 83 days of enactment
(By Feb. 4)

8. **Conclude negotiations** with contractor on modifications to the contract.

Within 85 days of enactment
(By Feb. 6, 1998)

9. **Prepare final modification**

Within 88 days of enactment
(By Feb. 9, 1998)

10. **All action completed, with NAGB and contractor sign-off**

Within 90 days of enactment
(By Feb. 11, 1998)

If contract modifications are made successfully, stop here. If the contract is to be terminated and a new contract negotiated, follow steps 11 – 15 below.

- | | |
|---|---|
| 11. <u>Institute contract termination procedures</u> | Within 90 days of enactment
(By Feb. 11, 1998) |
| 12. <u>Prepare NEW RFP</u> | Within 90 days of enactment
(By Feb. 11) |
| 13. <u>Publish RFP</u> in Commerce Business Daily | Within 105 days of enactment
(By Feb. 26) |
| 14. <u>Complete contract competition</u> | Within 151 days of enactment
(By April 13) |
| 15. <u>Make new award</u> | Within 165 days of enactment
(By April 28) |

prepared by: Ray Fields

Voluntary National Tests

Proposed Process for Review and Adoption of Test Specifications by the National Assessment Governing Board

Background and Overview

In his January 1997 State of the Union Address, President Clinton proposed a voluntary, annual reading test at grade 4 and a mathematics test at grade 8. These tests will provide parents and teachers with information about how their children are progressing compared to high standards and to students in other states, the nation, and various countries. An individual student's reading achievement level in fourth grade will be related directly to information obtained from the National Assessment of Educational Progress (NAEP) in reading. For math, a student's achievement in grade 8 will be linked to performance on the NAEP mathematics assessment, and to the Third International Mathematics and Science Study (TIMSS).

The Voluntary National Tests (VNT) will be based on NAEP Frameworks in reading and mathematics, developed by the National Assessment Governing Board (NAGB). The Voluntary National Test in Reading will be developed according to the 1992 - 1998 NAEP Reading Framework. For math, the VNT will be based on the 1996 NAEP Mathematics Framework. Through statistical procedures, student performance on the VNT in math and reading will be linked to the NAEP achievement levels—basic, proficient, and advanced—developed by the Governing Board.

To begin the VNT development process, the Department of Education contracted with Council of Chief State School Officers (CCSSO) and MPR Associates to convene panels of experts to recommend test and exercise specifications for reading and mathematics appropriate for individual testing. The groups completed their work on September 15, 1997. The Department also has awarded a contract to the American Institutes for Research (AIR) to prepare the test instruments, conduct linking studies to NAEP, and perform other VNT activities.

On September 25, Secretary of Education Richard Riley announced that most VNT development work would be suspended. He said he had decided not to forward officially CCSSO's recommended specifications to the test contractor (AIR), in order to give the Governing Board time "to ensure that the specifications for the national tests are fully consistent with those for NAEP." The Secretary said that only test specifications approved by the Board would guide development of the Voluntary National Tests.

On November 13, 1997 President Clinton signed into law P.L. 105-78, an appropriations bill giving the National Assessment Governing Board "exclusive authority over all policies, direction, and guidelines for developing voluntary national tests..." pursuant to the AIR contract. Under the law, the Board has 90 days after November 13 to review the AIR contract and, as the Board determines necessary, decide whether to modify or rescind it.

The purpose of this document is to propose a process for the Governing Board's review and adoption of VNT specifications, in accordance with P.L. 105-78 and the conference report language.

Process for Review of Test Specifications

The conference report for P.L. 105-78 summarizes the Act's content and provides some specific guidance not contained in the legislation. Of particular importance is a statement that gives the Governing Board exclusive authority over [AIR] "contract RJ97153001, which will be based on the same content and performance standards as are used for NAEP, and which are linked to NAEP to the maximum extent possible."

Therefore, in carrying out its responsibilities under this new legislation, the Governing Board will conduct a thorough review of the proposed specifications for grade 4 reading and grade 8 mathematics. The primary objective of this process is to prepare VNT specifications that are consistent with NAEP's content and achievement levels, and which yield valid and reliable individual test scores. The VNT specifications proposed by CCSSO would be reviewed with this goal in mind. The review will include a technical analysis of the proposed specifications, opportunity for public comment, and consideration of other analyses and reviews. Upon completion of this work, a set of recommended specifications for 4th grade reading and 8th grade mathematics will be presented for final Board action at its meeting of March 5 - 7, 1998.

Technical Analysis Contract

The Board should consider several major components in a specifications review process for the VNT. First, it is recommended that an external contractor be used to perform a comprehensive technical analysis of the proposed reading and math specifications to ensure consistency with NAEP frameworks. Given the critical importance and high visibility of this review process, the Board would best be served by working with a prestigious, independent, national policy or research organization. Such an organization should have no prior or current contractual relationship with the VNT development process. Further, the organization shall have a reputation for rigor, objectivity, and balance.

The statement of work for this short-term technical analysis contract would include, but not be limited to, a detailed description of sources of information to be used for the analysis, consultant qualifications, tasks, specifications review criteria, timelines, and products. The scope of work would clearly delineate the role of the Board, its committees, and its staff.

The main purpose of this contract would be to provide the Board with a detailed technical analysis of the proposed reading and math specifications. Major components of that work would include a thorough content analysis of the VNT specifications, items, and scoring criteria in relation to NAEP frameworks, specifications, items, and scoring criteria. The contractor shall identify all VNT deviations from NAEP, and classify those deviations that are and are not required to produce individual student test scores. In addition, the technical analysis should include a discussion of any non-standard test administration conditions recommended in the VNT specifications. The contractor also should report to the Board on differences between NAEP and VNT accommodations for students with disabilities and limited English proficient students.

Results of the technical analysis should be contained in a final report to the Board, to be delivered 45 days after the contract is awarded. Tasks, timelines, and dates for interim products, reports, and Board and staff review will be specified in the scope of work. The timeline for completion of this activity must be carefully coordinated with the Board's other major VNT responsibility—review of the test development contract.

Public Hearings

The conference report states that the National Assessment Governing Board shall hold public hearings on the VNT development activities. Therefore, a second component of the specifications review process shall involve such hearings to gather public comment. As stated in the conference report, the conferees intended that "the National Assessment Governing Board shall ensure that such hearings are widely publicized, and that activities conducted to publicize such hearings communicate effectively with the broad and diverse populations that may be affected by such tests."

These public hearings, sponsored and conducted by the Governing Board, would supplement the technical analysis of the specifications and allow the Board to gather comments from a broad range of groups and individuals. The hearings should include both invitational and open components, to permit maximum feedback from important VNT constituents and the general public.

The Board should conduct public hearings in early 1998 in two or three locations around the country. In preparation for the hearings, the Board should prepare a brief, clearly written summary of major NAEP features for grade 4 reading and grade 8 math, and the major points of the proposed VNT specifications. Any important differences should be highlighted. This document could serve as the focus of the hearings, to increase the likelihood that participants would provide relevant, useful comments to the Board.

A thorough analysis of the written testimony and public hearings transcripts will be provided to the Board to inform its deliberations and final action on the VNT specifications in March.

NAGB Staff Review

In addition to the technical analysis contract, NAGB staff will prepare an independent review of the proposed VNT specifications in relation to the NAEP reading and mathematics framework, specifications, and items, and recommend modifications or changes to the VNT materials, as necessary. Such a review will supplement the external analysis and the public hearing comments, to inform the Board's deliberations on the VNT specifications.

Opportunity for Written Comment

In keeping with its responsibilities in P.L. 105-78, the Board should publish a Federal Register notice inviting public comment on the specifications the Board should adopt for the VNTs in reading and mathematics. To reach the broadest possible audience, the Board should use additional means of publicizing this public comment period. A summary of these written comments will be prepared for Board consideration at the March 1998 meeting.

Use of the Internet

The Governing Board's Internet web site will be an integral component of the specifications review process. The web site will provide announcements of public hearings, permit on-line registration for those hearings, contain summary and complete versions of the proposed VNT specifications (including sample test items), allow on-line ordering of specifications documents, and provide users with an opportunity to submit comments via email on the specifications.

NAS Workshop

The National Academy of Sciences (NAS) plans to conduct a workshop in December 1997 on the current VNT specifications and their relationship to NAEP. Several Board members and staff should plan to attend this workshop. All workshop proceedings should be summarized and presented to the Board at its January 1998 meeting.

Proposed Timelines for VNT Specifications Review

Date	Activities
1997 November	Clarify committee assignments Develop and discuss process for specifications review Obtain Board approval of the process for specs review Develop statement of work for technical analysis contract
December	Award contract to conduct technical analysis Finalize activities and timelines for specifications review Prepare draft staff review of specifications
1998 January	Develop interim report for January Board meeting on the technical analysis contract and staff analysis Take action on acceptable aspects of specifications, if possible Finalize preparations for public hearings Review NAS conference proceedings at January Board meeting Conduct public hearings; summarize proceedings
February	Review final technical analysis report from contractor Convene Subject Area Committees to review specifications analyses from contractor and staff and other input Compile all materials for March Board meeting Prepare final recommendations on specifications
March	Board takes final action on VNT specifications



National Assessment Governing Board

National Assessment of Educational Progress

FAX TRANSMITTAL COVER SHEET

DATE: 8/12

TRANSMITTED TO: Mike Cohen

FAX NUMBER: 456-5581 PHONE #: _____

FROM (SENDER): Roy Truby

FAX #: 202-357-6945

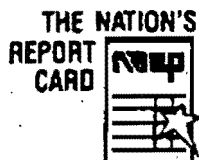
PHONE #: 202-357-6938

MESSAGE TO RECEIVING PARTY: _____

The usual the letter went to members
of the Outlaying & Appropriation Committee
Also we have had a request from the
Hispanic Education Coalition for the V.N.T
update letters. as I mentioned Bob Eckhardt asked
for a personal briefing on V.N.T. matters.

TOTAL NUMBER OF PAGES TRANSMITTED, INCLUDING COVER SHEET: _____

IF THERE ARE ANY PROBLEMS WITH THIS TRANSMISSION, PLEASE NOTIFY THE
SENDER IMMEDIATELY AT 202-357-6938.



National Assessment Governing Board

National Assessment of Educational Progress

August 10, 1998

The Honorable Edward M. Kennedy
Ranking Minority Member
Committee on Labor and Human Resources
U.S. Senate
Washington, DC 20510

Dear Senator Kennedy:

This is to provide a report on actions taken by the National Assessment Governing Board regarding proposed voluntary national tests at the Board's August 6-8 quarterly meeting. The law pertaining to the voluntary national tests, Public Law 105-78, gives the Governing Board exclusive authority over all policies, direction, and guidelines for developing voluntary national tests pursuant to contract RJ97153001, originally awarded by the Department of Education. The contractor is the American Institutes for Research.

The actions taken by the Governing Board include: (1) continuation of the test development contract for a second year; (2) plans for public hearings on inclusion and accommodations in the voluntary national tests for students with disabilities and students with limited English proficiency; and (3) policies for proposed pilot tests.

Guidance on whether to continue the test development contract

The Governing Board has neither endorsed nor opposed the initiative to develop voluntary national tests in 4th grade reading and in 8th grade mathematics. The Board views its charge under Public Law 105-78 as follows: to develop the best possible tests and carry out activities consistent with relevant authorizing and appropriations legislation enacted into law.

In accordance with its charge, the Governing Board changed the period of performance for the test development contract to end on September 30 each year rather than August 15, as the original contract stated. This date change was designed to permit the Board to take into account relevant authorizing or appropriations legislation for FY 1999 enacted into law during FY 1998. Such legislation has not yet been enacted into law.

Under the contract, the Board must provide the contractor with advance notice about continuation or termination. Specifically, the contract requires that the Governing Board provide the contractor by September 1 of each year with a non-binding notice of the Board's intent with respect to continuation of the contract for an additional 12 months. Negotiations on the contract must be concluded by September 30 of each year, should the decision be to continue the contract.

Accordingly, to set the contract review into motion, the Governing Board on August 7, 1998 unanimously passed a resolution (enclosed) outlining the process. The process is based on adherence to existing law. Following existing law and the contract requirements, the Governing Board resolution:

- (1) directs Governing Board staff to notify the contractor, by September 1, 1998, of the Board's intent to continue the contract through September 30, 1999;
- (2) authorizes the Executive Committee of the Governing Board to approve the decision letter on whether to continue the contract, which is due to the contractor on September 15, 1998 and which will be based on the adequacy of the contractor's response to the September 1 notification letter and relevant legislation enacted into law;
- (3) provides that negotiations must be completed by September 30, 1998, if the decision is to continue the contract; and
- (4) requires that the September 1 notification letter, the September 15 decision letter, and the contract extension, should there be one, include an explicit acknowledgement that the status of the contract must be consistent with relevant authorizing or appropriations legislation that may be enacted into law and that the contract may be revised or terminated accordingly.

Approval of plans for public hearings on inclusion and accommodations

The Governing Board approved plans to conduct hearings on inclusion and accommodations for students with disabilities and for students with limited English proficiency. Public law 105-78 requires the National Academy of Sciences to conduct a study on fairness and appropriate use of tests, due September 1, 1998, which will address inclusion and accommodation issues. The conference report for PL 105-78 asks the Governing Board to conduct public hearings on the NAS reports. The Board's public hearings on inclusion and accommodations will be based on the recommendations in the NAS report.

The public hearings will be conducted in October and November at six sites across the country. Proposed sites are Atlanta, Austin, Chicago, Los Angeles, New York City, and Washington, DC. At each site, there will be a separate ½-day hearing on inclusion and accommodations for students with disabilities and for students with limited English proficiency. There will be a transcript and a written summary for each ½-day hearing. In addition, two reports will be prepared (one addressing students with disabilities, the other addressing students with limited English proficiency), covering all six hearings and any additional written testimony that may be submitted.

The results of the public hearings will be published and entered on the National Assessment Governing Board web site. In addition, the results will be used by the Governing Board to inform policy development related to inclusion and accommodations for the proposed field test for the voluntary national test, scheduled to be conducted in March 2000 unless prohibited by federal law.

Policies for proposed pilot tests

The pilot tests for 4th grade reading and 8th grade mathematics test questions are scheduled to be conducted in March 1999, unless prohibited by federal law. The purpose is to determine whether the individual test questions are good, the level of difficulty of each question, and whether any question should be eliminated because of race, gender, or cultural bias. Based on the results, decisions will be made about whether to keep or reject each test question and about how to construct test forms for the field test. Because the pilot test focuses on how individual items function, no student or school will receive test scores from the pilot.

The Governing Board gave conditional approval to three proposed policies that are limited to the pilot tests. These policies address: (1) data collection for the pilot test; (2) data analysis for the pilot test; and (3) inclusion and accommodations for the pilot test. The Board directed that several changes be made in the policies to clarify the meaning of certain terms and to make minor editorial improvements.

Thank you for your interest. Please feel free to contact me if you have any questions.

Sincerely,



Mark Musick
Chairman

Enclosure



National Assessment Governing Board

National Assessment of Educational Progress

August 11, 1998

The Honorable Bob Etheridge
U.S. House of Representatives
Washington, DC 20515

Dear Representative Etheridge:

As you requested, I am enclosing a set of materials about the voluntary national test initiative. I have included reports about decisions made by the National Assessment Governing Board previously sent to members of cognizant authorizing and appropriations committees of Congress. Also, having added your name to our list of Members to receive these reports, I am also including the report from the August 6-8, 1998 meeting.

One of several significant actions taken by the Governing Board pertains to the test development schedule. Under the original Department of Education contract, the test developer was to develop and pilot test questions by September of 1997 and field test proposed test booklets in March of 1998. This was needed to achieve the goal of administering actual tests in March of 1999.

The Governing Board has revised the schedule. The current schedule calls for development of test questions during 1998, pilot testing of the questions in March of 1999, and field testing of proposed test booklets in March of 2000. This schedule would result in actual testing of students in March of 2001.

It is important to understand the role that pilot testing plays in test development.

The purpose of the pilot test is to determine whether the individual test questions are good, the level of difficulty of each question, and whether any question should be eliminated because of race, gender, or cultural bias. As a result of the pilot tests, decisions will be made about whether to keep or reject each test question and about which questions should go on which proposed test booklets in order to build test forms for the field test that are equivalent. Because the pilot test focuses on how individual items function, no student or school will receive test scores from the pilot.

Thank you for your interest. Please feel free to contact me or Ray Fields of my staff if you have any questions.

Sincerely,

Roy Truby
Executive Director

Enclosure

800 North Capitol Street, N.W.
Suite 825, Mailstop 7583
Washington, D.C. 20002-4233
Phone: (202) 357-6938
Fax: (202) 357-6945