DOCUMENT RESUME

ED 360 397 TM 020 653

AUTHOR Cizek, Gregory J.
TITLE Reactions to National Academy of Education Report, "Setting Performance Standards for Student Achievement."
SPONS AGENCY National Assessment Governing Board, Washington, DC.
PUB DATE Aug 93
NOTE 41p.; Paper commissioned by the National Assessment Governing Board.
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Academic Standards; Achievement Tests; Educational Policy; Elementary Secondary Education; *Evaluation Methods; Literature Reviews; National Competency Tests; *Performance; Psychometrics; Research Problems; *Research Reports; *Test Reliability; Test Validity
IDENTIFIERS Angoff Methods; National Academy of Education; National Assessment Governing Board; *National Assessment of Educational Progress; *Standard Setting; Trial State Assessment (NAEP)

ABSTRACT

The report of the National Academy of Education (NAE), "Setting Performance Standards for Student Achievement," provides an interpretation of NAE investigations into the procedures surrounding establishment of achievement levels for the National Assessment of Educational Progress (NAEP). The NAE report synthesizes the investigations of other research efforts commissioned by the NAE Panel on the Evaluation of the NAEP Trial State Assessment. The following documents have been reviewed: (1) the NAE report itself, with 10 studies commissioned in preparing the report; (2) "Setting Achievement Levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing" from the American College Testing Program, 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" from American College Testing, 1993. It is concluded that the NAE evaluation is a seriously inaccurate representation of the technical and procedural propriety of the NAEP levels-setting process. Errors and inaccuracies are described in the general areas of authority and credibility, conclusions about the use of the Angoff method, procedural propriety and results, and the validity of NAEP levels. Accurate conclusions are also reviewed, and some constructive suggestions are made. (Contains 35 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Reactions to National Academy of Education Report,

"Setting Performance Standards for Student Achievement"

ED 363 957

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

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

August 1993

BEST COPY AVAILABLE

IR020653

ERIC

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 th 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 difficult 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.