DOCUMENT RESUME

ED 360 398 TM 020 654

AUTHOR Kane, Michael
TITLE Comments on the NAE Evaluation of the NAGB Achievement Levels.
SPONS AGENCY National Assessment Governing Board, Washington, DC.
PUB DATE Sep 93
NOTE 20p.; Paper commissioned by the National Assessment Governing Board.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Academic Standards; Achievement Tests; Educational Policy; Elementary Secondary Education; *Evaluation Methods; Literature Reviews; Measurement Techniques; 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 National Academy of Education (NAE) Panel has drawn two major conclusions in its evaluation of the National Assessment Governing Board's (NAGB) efforts to set achievement levels for the National Assessment of Educational Progress (NAEP). They are that the Angoff procedure is fundamentally flawed for the setting of achievement levels, and that the weight of evidence suggests that the 1992 achievement levels were set unreasonably high. However, the evidence presented in the final report of the Panel and in the reports of the studies it commissioned do not justify these conclusions, and some of the evidence directly contradicts them. The NAE report is well written and provides a good discussion of the general issues in standard setting, but it generally accepts the results of the studies it commissioned at face value, without regard to their flaws, and it 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 it proposes for reporting the NAEP results. Nine commissioned studies are evaluated individually. (SLD)

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 & De Stefano, 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 IAEP international Results. **Beaton A.E., & Gonzalez, E.J., August, 1993**

By relating results on NAEP and the IAEP, 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 IAEP.