The Validity and Credibility of the Achievement Levels for the 1990 National Assessment of Educational Progress in Mathematics

CSE Technical Report 330

Robert L. Linn
The University of Colorado

Daniel M. Koretz
The Rand Corporation

Eva L. Baker and Leigh Burstein
UCLA
Center for Research on Evaluation, Standards
and Student Testing

With contributions from a subgroup of the Technical Review Panel for Studies of the Validity of the National Assessment of Educational Progress.

The project reported herein was performed under a contract from the National Center for Educational Statistics (NCES). However the Opinions expressed herein do not necessarily reflect the position or policy of NCES and no official endorsement by NCES should be inferred. The report was completed in January, 1991, but held for release until after the release of the Trial State Assessment results on June 6, 1991. The National Assessment Governing Board collected new data following the completion of this report that provide the basis for their achievement levels.

The Validity and Credibility of the Achievement Levels for the 1990 National Assessment of Educational Progress in Mathematics

SUMMARY

Background and Recommendation

The statute currently authorizing the National Assessment of Educational Progress (NAEP) calls for the National Assessment Governing Board (NAGB) to set appropriate achievement levels in all areas and grades tested by the NAEP. These achievement levels are intended to indicate what students **should** know, not merely what they currently **do** know. During the last half of 1990, NAGB undertook a major first step in this direction and conducted a project to set achievement levels for the 1990 mathematics assessment. Three achievement levels were posited (advanced, proficient, and basic) at each grade level (4, 8, and 12). At issue is whether these achievement levels should be used for reporting the national or state-level results of the NAEP.

This preliminary report, part of a Congressionally mandated study of the validity of the NAEP, examines the reasonableness and technical adequacy of the achievement levels. Our study focuses on the results of the NAGB effort, not on the arguments for and against the idea of achievement levels or the adequacy of the procedures used.* Our study addresses a fundamental question: Are the achievement levels adequate for supporting the conclusions or inferences for which they will be used?

Our analyses indicate that the achievement levels are seriously flawed--seriously enough that they cannot credibly support the conclusions to be based upon them. Moreover, we do not believe that practical, **post hoc** adjustments can remedy the problems we have found. Because the flaws in the achievement levels could undermine the credibility, not only of the achievement levels, but of NAEP itself, **we recommend that the achievement levels developed so far not be used in any public reporting of national or state-level NAEP results.**

---

* The process used to set achievement levels was evaluated in a separate study commissioned by NAGB and conducted by Daniel Stufflebeam, Richard Jaeger, and Michael Scriven.

Bases for Our Recommendations

We judged the achievement levels both in terms of conventional criteria of validity and reliability and in terms of NAGB's explicit goals for the achievement levels, which set several criteria for their evaluation. NAGB's stated goals indicate that the achievement levels should reflect consensus judgements. In addition, they should be grade-specific standards, representing levels of mastery of material presented by the grades in question.

One critical criterion for evaluating the achievement levels is their robustness--that is, their coherence and consistency across groups of raters. The importance of consistency across groups of raters is underscored by the goal that the achievement levels should represent a consensus, but it would be essential regardless. Policymakers and the public need to be confident that the proportion of students reaching the achievement levels tells them something about the student population rather than about the particular group of judges used for a given grade and subject.

In a substantial number of instances, the differences among groups of judges participating in the NAGB effort were so large as to undermine the credibility of the results. That is, seemingly arbitrary choices among the panel of judges would lead to fundamentally different conclusions about the condition of student achievement. In the worst case, the 11 judges at grade 4 who attended only the initial achievement level meeting in Vermont would have set the basic achievement level 14 points higher on the NAEP scale than the 11 judges at grade 4 who returned to the second meeting in Washington. A difference of that magnitude could correspond to change the percentage of students achieving that level from roughly 50% to 65%.

Another important aspect of robustness is the coherence of the ratings across the three grades rated (4, 8, and 12). Achievement goals should demand more of students in higher grades. In a number of respects, the ratings were not coherent across grade levels. That is, judges set achievement levels that would require 8th grade students to perform as well or better than 12th grade students. This finding appears in the ratings for test items given to students at both grades and occurs despite considerable evidence that 12th grade students perform substantially higher than 8th grade students on those common items.

A major goal of the setting of achievement levels is to shift attention to what students should know rather than simply what they do know. In the language of testing, the current NAEP reporting is normative; following the identification of achievement levels, the reporting would be criterion-based. The evidence available to us does not directly address the question of whether the achievement levels are in fact criterion-based. We

approached the issue indirectly, however, and posed the following question: do the achievement levels show the patterns one might expect if they actually were criterion-based?

In general, the data suggest strongly that the achievement levels are not truly criterion-based and therefore cannot support the judgmental conclusions which provide their rationale. This conclusion is based on the empirical relationships between the achievement level ratings and the actual normative performance of students. Those relationships are generally very high (with correlations often over .90); vary unexpectedly little across content areas (such as geometry and numbers and operations); and do not differ across different process categories (concepts, procedures, and problem solving).

An additional criterion against which the achievement levels must be evaluated is the precision of measurement available in NAEP for students performing close to the established achievement levels. As other observers (e.g., Anrig, 1991) have noted, this is likely to be a problem primarily at high levels of achievement. Our analysis confirms that the NAEP mathematics assessment, as currently designed, is not aimed at a high enough level to provide good measurement at some of the suggested achievement levels, particularly the advanced level in grades 4 and 8.

Conclusion

The findings reported here indicate that on several different grounds, the achievement levels established at this point are not credible and therefore cannot be useful for interpreting educational performance. They do not reflect consensus judgement, vary markedly from one group of raters to another, are incoherent across grade levels, appear not to be truly criterion-based, and overreach the capabilities of the test in terms of measurement precision.

In addition, we should point out that there is as yet no evidence supporting one of the key assertions made about the achievement levels--that is, that the proficient levels represent adequate preparation for subsequent schooling. The findings reported here, while not intended to test that claim, do offer reasons to doubt its reasonableness. Unless one believes that students need no more preparation to succeed as freshmen in college than in grade 9, the proficient level should be substantially higher in grade 12 than in grade 8. As noted, the current achievement levels require equal or higher levels of performance in grade 8 than in grade 12, where test items were rated for both grades.

Given these problems, the use of the current achievement levels for reporting national or state-level NAEP results would be a serious error. It would lead the public and policymakers to make insupportable inferences about student performance, and this in turn could lead to undesirable and possibly counterproductive

changes in policy. Moreover, **using achievement levels that suffer from such severe flaws could undermine the credibility, not only of the achievement-levels process, but of the NAEP itself.** The public and policymakers, in turning to the NAEP for invaluable information about student performance, may fail to distinguish between the flaws of the achievement levels and the strengths of NAEP.

To put these pessimistic conclusions in perspective, we note that the task that NAGB undertook was immense and in many respects unprecedented. We know of no previous effort to establish achievement levels that would serve as "appropriate achievement goals" for the nation as a whole. Moreover, the effort had to be undertaken on a extremely tight time schedule using existing instruments rather than ones designed with the achievement levels in mind. We commend NAGB and the many dedicated professionals who participated in this difficult undertaking. While this first effort did not yield achievement levels that can be used for reporting NAEP results, it did provide a great deal of information that should prove invaluable in future efforts of this sort.

There are a number of fundamental issues worthy of careful consideration if new efforts to set achievement levels are undertaken for 1992 or beyond. The definition of the population of judges, possible stratifications of judges, and sampling from those strata will require detailed consideration. The basic approach also deserves further consideration. The approach of rating individual items in the existing item pool rests on a questionable assumption that complex determinations of student proficiency can be inferred from these ratings. It may be desirable to consider alternative approaches that do not rely exclusively on the aggregation of ratings of individual items. Whatever method is undertaken next, it should provide evidence of robustness, coherence, and the criterion basis of the achievement levels.

THE VALIDITY AND CREDIBILITY OF THE ACHIEVEMENT LEVELS FOR THE 1990 NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS IN MATHEMATICS

During the last half of 1990, the National Assessment Governing Board (NAGB) undertook a major effort to establish achievement levels for the 1990 mathematics assessment of the National Assessment of Educational Progress (NAEP). The results of that work to set achievement levels are the focus of this report. More specifically, the purpose of the present report is to provide an evaluation of the adequacy of the achievement levels. The emphasis is on the results, not the process of obtaining judgments. The process was evaluated in a separate study, commissioned by NAGB and conducted by Daniel Stufflebeam, Richard Jaeger, and Michael Scriven.

Five major issues guided our analyses. First, we focused on the issue of **robustness of the achievement levels**. Subsumed under this general issue are questions about the degree to which the achievement levels set for the three grades are coherent and the extent to which the results are consistent across different groups of raters within a grade. Second, we focused on the **interpretive basis for the achievement levels**. More specifically, we considered the potential claim that the achievement levels are criterion-referenced rather than normative standards. Third, we analyzed the **precision of measurement** of students performing at the established achievement levels with existing NAEP item sets. Fourth, we considered problems with the **process and the use of adjustments**. Finally, we addressed the broader questions related to the **validity of the interpretations of results based on the achievement levels, or, more generally, asked whether the interpretations are credible**.

Our analyses indicate that **the achievement levels are seriously flawed** -- seriously enough that they cannot credibly support the conclusions that will be based upon them. Accordingly, **we recommend that the standards developed so far not be used in any public reporting of national or state-level NAEP results**.

The detailed analyses that led to the above conclusion and recommendation are provided in the section of the report starting on page 5 under the heading "Evaluation". To place those analyses in context, however, we begin with a brief description of the background for the project and the procedures used to set the achievement levels.

BACKGROUND

NAGB Policy Framework and Technical Procedures.

Reasons for the NAGB decision to undertake the project to set achievement levels were clearly stated in the Policy Framework and Technical Procedures adopted by NAGB on May 11, 1990.

> "Among the most significant responsibilities of the National Assessment Governing Board are (1) 'taking appropriate actions ... to improve the form and use of the National Assessment' and (2) setting 'appropriate achievement goals' for each grade and subject tested under the NAEP. The two responsibilities fit well together. By defining the levels of appropriate achievement on the National Assessment the Board will increase greatly the significance and usefulness of NAEP results to educators, policymakers, and the American public" (NAGB, May 11, 1990, p. 1).

The Policy Framework adopted by NAGB went on to discuss the need for, and potential benefits of, established achievement levels against which student performance can be compared. Three achievement levels were proposed for each grade and subject assessed by NAEP. These three levels were referred to as Basic, Proficient, and Advanced.

The Proficient level was expected to "represent solid academic performance for each grade tested--4, 8, and 12--and reflect a consensus that students reaching such a level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling" (NAGB, May 11, 1990, p. 5). The Advanced and Basic levels, as the labels suggest, were expected to denote achievement levels that were respectively higher and lower than the Proficient level. More specifically, the Advanced level was described as "superior performance beyond proficient grade-level mastery at grades 4, 8, and 12" while the Basic level was depicted as "partial mastery of the knowledge and skills that are fundamental for proficient work at each grade" (NAGB, May 11, 1990, p. 5).

In addition to approving the Policy Framework for the setting of achievement levels, NAGB approved the Technical Procedures to be used in setting the achievement levels (May, 11, 1990). The Technical Procedures included identification of the judgement procedure (the "modified Angoff procedure"); a description of the composition, size, and training of the judges; and a charge that the Board "conduct a formal evaluation of the process" (NAGB, May 11, 1990, p. 21).

The Technical Procedures provided a guide for the conduct of the achievement setting effort under the leadership of Professor Ronald Hambleton. The charge for a formal evaluation led to the commissioning of a comprehensive evaluation by a three person team (Professors Daniel Stufflebeam, Richard Jaeger, and Michael Scriven) which was headed by Professor Stufflebeam. The present analysis of the achievement levels is independent of the work of both Hambleton and the Stufflebeam, et al., evaluation team. Our analyses depend heavily, however, on data and results obtained from Hambleton.

Vermont Meeting of Panels of Judges.

A meeting of 63 judges (22 for each of grades 4 and 8, and 19 for grade 12) was convened in Essex Junction, Vermont on August 16 and 17, 1990. Prior to the meeting, the judges were provided with background materials including the 1990 Mathematics Objectives, a NAGB Policy Framework and Technical Procedures report, and sample NAEP mathematics items. The judges were given a Handbook for Judges (Hambleton, August, 1990) which included definitions of achievement levels and described the process to be used for rating items. Professor Hambleton conducted a training session in the use of the modified Angoff procedure at the beginning of the meeting.

At each grade level judges were organized into four smaller groups of 5 or 6. After a discussion of the definitions of the three achievement levels judges provided their initial ratings for each item (round 1). Judges were then given normative student performance data consisting of the proportion of students who answered each item correctly (p-values) and plots of item-test regressions and then completed a second round of item-by-item judgments.

Finally, a third round of judgments was obtained which, according to the original plan, was designed to follow group discussion of item ratings from the first two rounds. In particular, the high and low ratings of each item at each achievement level were to be "identified and reasons discussed for these ratings, along with pertinent points about the item" (Hambleton, August, 1990, p. 11). After discussion of an individual item, judges completed their round three ratings and then moved to the next item. As was noted by Stufflebeam, et al. (1991), however, a substantial number of the raters reported that there was insufficient time for the task.

Washington Meeting of Panels of Judges.

Due to concerns about time limitations and issues raised by NAGB's evaluation team about some aspects of the Vermont results, a second meeting was held to continue the process of obtaining ratings that could be used to establish the achievement levels. Thirty eight of the 63 judges who were in Vermont were re-convened in Washington, DC on September 29 and 30, 1990. The breakdown by grade level of the returning judges was 11 of 22 for grade 4, 19 of 22 for grade 8, and 9 of 19 for grade 12.

Before asking judges to provide a fourth round of ratings, a two-hour discussion of the definitions of the three achievement levels was led by Professor Hambleton. For the fourth round of ratings, judges were instructed to consider only the difficulty of items for students who were considered marginal for each achievement level. In particular, judges were instructed to ignore questions of appropriateness of the items in making their achievement level ratings.

After completing the fourth round of ratings the judges were presented with results based on the ratings from all four rounds. The differences in round three results for the four separate groups of judges at each grade level were highlighted. Also highlighted were comparisons of ratings of items common to more than one grade. The latter results called attention to the fact that a higher standard was required at grade 8 than at grade 12 for most items common to those two grades.

Meetings of judges held separately according to grade level being rated and combined across grade levels were held to consider the results obtained through round 4. The importance of consistency and coherence of the achievement levels was emphasized in those meetings.

Finally, the fifth and last round of ratings was obtained. The round 5 ratings differed from rounds 1 through 4 in that each judge was asked to provide a single rating of the average percent correct expected for each of the three specific achievement levels. That is, individual items were not rated in round 5. Specifically, judges were asked to give three percents (Basic, Proficient, and Advanced) in response to the following instruction: "On the basis of (1) my personal item ratings, (2) discussions with members of my work group, other participants at the same grade level as myself, and participants at the other two grade levels, and (3) the statistical data I had an opportunity to review, my recommended marginal achievement levels are as follows: Basic _____% Proficient _____% Advanced _____%."

Analyses, Adjustments and Recommendations from Hambleton and Technical Advisory Committee on Standard Setting.

A number of analyses of the rating data were conducted by Hambleton. Advice regarding some of these analyses was provided by a Technical Advisory Committee on Standard Setting (TACSS) consisting of Professors Robert Forsyth, Edward Haertel, and Richard Jaeger. The details of those analyses are summarized in a memorandum from Hambleton to Roy Truby, NAGB Executive Director, dated December 18, 1990.

Several adjustments in the round 5 ratings were considered and two were recommended for use. Briefly the suggested adjustments, rationales, and recommendations were:

(1) Two categories of items that were included in the Vermont ratings were recommended for exclusion from the achievement setting process. These items are referred to as higher order thinking skills and estimation items. Although both types of items are part of national NAEP, they are not part of the booklets used at grade 8 for the trial state assessment. Hambleton (December 18, 1990) recommended with the concurrence of the TACSS that these items be excluded when ratings are combined to set achievement levels.

(2) Due to the skew in the distribution of ratings for some of the achievement levels, it was recommended by Hambleton (December 18, 1990), with the concurrence of the TACSS, that medians rather than means be used to establish achievement levels.

(3) Consideration was given to adjusting round 5 ratings for differences in ratings at round 3. This was for judges who were unable to attend the Washington meeting compared to those who attended. As is discussed in greater detail below, Hambleton's (December 18, 1990) analyses indicated that such adjustments would have sizeable effects on achievement levels only at grade 4. Because of differences in process at the Washington meeting in comparison to the Vermont meeting and lack of strong theoretical basis for an adjustment procedure, it was recommended by Hambleton and the TACSS that no adjustments be made for the change in sample of judges from Vermont to Washington.

(4) The TACSS suggested "smoothing" the achievement levels to enhance consistency and coherence (see Hambleton and Jones, December 7, 1990). But specific smoothing procedures were not articulated and no action was taken on this suggestion.

(5) It was suggested that standard errors of the proficiency scale points corresponding to achievement levels be obtained based on the ratings (see, Haertel's notes to Mary Lyn Bourque dated November 18, 1990). The standard errors would provide a means of demonstrating the magnitude of change in percentage of students in each category defined by the achievement levels that might be attributed to uncertainty in the location of achievement levels.

EVALUATION

Before discussing the specific issues addressed in our analyses it is important to provide some context regarding the nature of the challenge that NAGB faced and the magnitude of the effort. Although procedures similar to those used by the panel of judges assembled by NAGB have been used in other situations to set passing scores for minimal competency tests, for certification examinations, or other types of tests where pass-fail decisions are made about individuals, we know of no previous experience that is comparable to the task that NAGB confronted, that is, the establishment of achievement levels that would serve as "appropriate achievement goals" for the nation. In addition, the achievement levels were expected to provide an improved means of reporting current achievement and for monitoring trends in achievement in the future.

The task was not only new, but the effort had to be undertaken on a short time-line and relied on existing assessment instruments rather than designing the assessment instruments around the desired achievement levels. We commend NAGB and the many dedicated professionals who participated on the panel and

worked on the design of the procedures and analyses of results. The effort was indeed extraordinary. Regardless of decisions that may be reached about the use or potential utility of the achievement levels, it is clear that much was learned that should well serve future work designed to establish appropriate achievement goals for the nation.

With this context in mind, we turn to our analyses of the five major issues identified in the introduction, beginning with the issue of the robustness of the achievement levels.

ROBUSTNESS OF THE ACHIEVEMENT LEVELS

Are the achievement levels sufficiently robust to be useable? This question guided our initial analyses of the rating data. Robustness of ratings, i.e. their consistency across groups of raters, provides one critical criterion for evaluating the achievement levels. The importance of consistency across groups of raters is underscored by the stated goal that the achievement levels should represent a consensus, but it would be essential regardless. Policymakers and the public need to be confident that the proportion of students reaching the achievement levels tells them something about the performance of the population of students rather than about the particular group of judges used to set achievement levels for a given grade.

To answer the robustness question we focused on three more specific questions: (1) Are the results coherent across grades? (2) To what extent are the results dependent on the specific composition of the raters who were able to attend the second rating session in Washington? (3) How consistent are the results from one group of raters to another within each grade?

1. Coherence of Levels from Grade to Grade.

A major issue that attracted considerable attention following the initial results from Vermont is the coherence of the achievement levels from grade to grade. As was previously noted, some NAEP items are used at more than one grade level. Among the items that were rated in Vermont, 32 items were common to grades 4, 8, and 12, and independent ratings were obtained from the three grade-specific panels of judges. Twenty seven additional items were common to grades 4 and 8 only and 39 additional items were common to grades 8 and 12 only. The later two sets of common items received independent ratings from the two grade-level appropriate panels.

Comparisons of ratings for the common items at the different grade levels raised questions about the coherence of the achievement levels for grades 8 and 12. Table 1, which is based on Table 11 from Hambleton and Jones (December 7, 1990), lists the average ratings from round three for the 32 items common to grades 4, 8, and 12, and the 39 items common to grades 8 and 12

only. Also shown are the average proportions of students at grades 8 and 12 who answered those items correctly. As can be seen, **the Proficient and Advanced level ratings were equal or nearly equal at the two grade levels. Basic level ratings were actually higher for both sets of common items at grade 8 than at grade 12, despite the fact that the average proportion correct for grade 12 students is substantially higher than the corresponding proportion correct for grade 8 students.** (See, also Figures 2 and 4 for a graphical display of the results in Table 1.)

The frequency distribution of the differences between the Basic ratings for grade 12 and those for grade 8 is shown in Figure 1. The 32 items common to grades 4, 8, and 12, and the 39 items common to grades 8 and 12 only, are combined in Figure 1. As can be seen, 10 of the 71 items rated for both grades 8 and 12 had essentially equal ratings (differences between -2% and +2%). Only 6 of the items were rated higher by 3% or more for grade 12 than for grade 8 while the remaining 55 items were rated higher by at least that amount for grade 8 students than for grade 12 students. For 12 of the items the grade 8 ratings exceeded the grade 12 ratings by 12% or more.

It should be noted that the definition of the 12th grade Basic level used in Vermont noted that "this level will be higher than minimum competency skills (which normally would be taught in elementary and junior high schools) and will cover significant elements of standard high school-level work." Such a definition together with the increase in the average proportion of students answering common items correctly from grades 8 to 12 makes the reversal of the grade 8 and 12 ratings appear all the more unreasonable.

When all grade 8 items and all grade 12 items are rated, the resulting average Basic achievement levels for round 3 are 70.6 at grade 8 and 54.2 at grade 12 (Hambleton & Jones, December 7, 1990, Tables 2 and 3). These ratings in the required percent correct metric are not directly comparable because different, albeit partially overlapping, sets or items are rated at the two grades. The Basic levels can be made comparable by converting the percent figures to the NAEP scale.

The empirical test characteristic curves for the composites at grade 8 and 12 (provided to NAGB by ETS on November 8, 1990) were used to convert the above average round 3 ratings at grades 8 and 12 to the NAEP scale score. The resulting Basic levels are 292 at grade 8 and 294 at grade 12. Although the conversion of round 3 ratings of all items to the NAEP scale using the empirical test characteristic curve for the composite yields a slightly lower Basic level at grade 8 than at grade 12, the two values (292 and 294) are disturbingly close together.

Regardless of how closs the two implied achievement levels are, it may seem surprising that the reversal that was found for

common items is not present when all items are rated and results are converted to the NAEP scale. There are two reasons for the disappearance of the reversal. First, items that are unique to grade 12 are generally more difficult than items unique to grade 8. Second, as will be shown below, the regression of ratings on empirical item difficulties has a slope that is substantially less than 1.0. **Consequently, the difficult items that are unique to grade 12 tend to have Basic ratings that are higher than the empirical item difficulties whereas the relatively easier items unique to grade 8 tend to have basic ratings that are lower than the empirical item difficulties.**

Round 4 results may also be used to make comparisons of the grade 8 and grade 12 ratings of items common to grades 8 and 12. Despite the fact that the round 4 ratings are based on fewer judges than the round 3 ratings, these comparisons have four advantages over the ones in Table 1. First they are limited to the reduced item sets, i.e., the so-called higher-order thinking skills and estimation items that were excluded from the analyses used to establish achievement levels, were also excluded from the comparisons. Second, the round 4 ratings were the final ratings obtained at the individual item level and are based on the subset of judges used to provide overall ratings. Third, it is the round 4 ratings of the reduced item sets versus the overall pool that were used to make adjustments in the round 5 global ratings in an attempt to account for the exclusion of higher-order thinking skills and estimation items. Finally, the judges in Washington had the benefit of additional training and discussion of the issue of coherence across grade levels.

The average p-values for 8th and 12th grade students on the reduced set of common items are shown in Table 2. Also shown for these common items are the average achievement level ratings from round 4, and the number of items in the reduced item set that are common to all three grades and common to grades 8 and 12 only. The exclusion of higher-order thinking skills and estimation items reduces the number of items common to grades 4, 8, and 12 from 32 to 21 and reduces the number common to grades 8 and 12 only from 39 to 19. As was true in Table 1 for the full set of common items, the average proportion correct was higher for 12th grade students than for 8th grade students for items common to both grades.

For the 21 items common to all three grades, the round 4 average Basic ratings are 1 point higher at grade 12 than at grade 8. On the other hand, the Proficient and Advanced average round 4 ratings for those items are higher at grade 8 than at grade 12 (by 2 and 5 points, respectively). Moreover, the ratings of the 19 items common to grades 8 and 12 only, are higher for grade 8 than grade 12 by either 4 or 5 points at all three achievement levels. Thus, the reversal noted for round 3 ratings for the total item pool is not unique to those ratings. If anything, there are more

grade reversals in the round 4 ratings than in the round 3 ratings.

In our judgment, the higher ratings of common items at grade 8 than at grade 12 poses a fundamental problem for the credibility of the achievement levels. The fact that the 12th grade achievement levels obtained on the NAEP scale from items unique to the 12th grade, as well as those common to other grades, is higher than that obtained in a parallel fashion for the 8th grade, does not solve the problem identified by the results in Figures 2 through 5.

2. Changes in Composition of Panel from Vermont to Washington.

Another aspect of the robustness issue of the achievement levels concerns the degree to which comparable levels would be set by different groups of judges. Confidence in the achievement levels would be enhanced to the degree that the different groups of judges provide comparable achievement levels. On the other hand, large fluctuations from group to group would suggest that the particular results depend in an arbitrary way on the particular group of individuals that provides the ratings.

The previous section compared results for groups formed according to the grade level for which achievement levels were being set. There are several other relevant ways of defining groups for the NAEP achievement level effort. Here, we focus on the difference between ratings provided by judges who were present in Vermont only, and those who returned to the second session in Washington where the final two rounds of ratings were obtained.

Hambleton and Jones (December 7, 1990) document differences in ratings from judges present in Vermont only and those who were also present in Washington. Additional analyses of the composition of the two groups and possible implications of changes in the group of judges are also provided by Hambleton (December 18, 1990 memorandum to Truby).

Hambleton focused on grades 4 and 12 because 19 of the 22 grade 8 judges who were in Vermont attended the second rating meeting in Washington. As Hambleton's results show there was a greater return rate at both grades 4 and 12 for educators than for non-educators. At grade 4, judges present in Vermont only set significantly higher achievement levels than judges present in Washington.

We converted the average ratings obtained by round for judges in attendance and judges not in attendance at the Washington meeting to NAEP scale scores using the empirical test characteristic curves provided by ETS (November 8, 1990). The mean Basic achievement levels that correspond to the ratings in each round for those present in Vermont only and those present in Washington are presented in Table 3. Graphical displays of the

results listed in Table 3 are shown in Figure 6 for all three grades. Since the grade 8 and 12 graphs are so similar through round 3, those results are displayed alone using a larger scale in Figure 7 and the grade 4 results are displayed alone in Figure 8.

As can be seen in Figure 6, and is even more apparent in Figure 8 where the grade 4 results are displayed alone, a higher Basic achievement level is suggested by the 11 judges who were present in Vermont only than by the other 11 judges who were also present in Washington and completed the last two rounds of ratings. The higher Basic achievement levels at grade 4 for raters present in Vermont only, than for raters attending the Washington meeting that Hambleton noted, translates to a difference of 14 points on the NAEP scale.

Using previous NAEP mathematics results as a rough guide, 14 points would correspond to about a half of a standard deviation for grade 4 students. A difference of that magnitude could obviously have a major impact on the proportion of students who performed above the Basic level at grade 4. Although these conversions apply to the full item pools rather than the reduced item pool, the size of the difference for the reduced item pool is similar in the percent correct metric to that reported for the total item pool (see Hambleton and Jones, December 7, 1990, Table 20 and Hambleton, December 18, 1990, page 6). Hence, the relatively large difference in scale scores as a function of the raters attending the two meetings is not due to choice of the item pool.

The plot of the Basic achievement levels in Figure 6 makes other characteristics of the results quite apparent. As expected, the Basic achievement level for grades 8 and 12 is substantially higher than the Basic achievement level at grade 4 at all 5 rounds of ratings for both those attending only the Vermont meeting and those attending the Washington meeting. Differences between the grade 8 and 12 Basic levels do not emerge until rounds 4 and 5, however.

Although the grade 8 and 12 differences between the Basic achievement levels, set by those present at Vermont only and those present at Washington, are small, they are in the same direction as the grade 4 differences (see Figure 7). That is, the Vermont only Basic achievement levels are slightly higher than the Washington Basic achievement levels through round 3. The other notable feature of Figure 7 is that the grade 8 and grade 12 judges present in Washington set the same achievement levels at round 3, but the levels separated at round 4. One can't but wonder if the grade 12 judges felt pressured to set higher achievement levels in round 4 while the grade 8 judges felt pressured to set lower ones in round 4 due to the incoherence of the grade-to-grade results obtained in Vermont.

We conclude that the differences in the Vermont only and Washington ratings at grade 4, which Hambleton previously

identified as statistically significant, are too large to be ignored. Eleven judges is a small number to set national achievement levels for grade 4 students under the best of circumstances. That number appears all the more inadequate in the face of evidence that those judges provide systematically different ratings than the presumably equally appropriate set of 11 judges who were present in Vermont only.

3. Group Differences in Average Ratings.

One additional approach to the investigation of the robustness of the achievement levels is based on a comparison of groups that worked separately at each grade. As was previously, indicated, at the Vermont meeting the judges for a given grade worked in 4 separate groups. The round 3 results for the individual groups provide additional evidence regarding the sensitivity of achievement levels to the membership of the groups that do the ratings. Hambleton and Jones (December 7, 1990) provide means and standard deviations of the Basic, Proficient and Advanced achievement levels from the round 3 ratings separately by group in Tables 4, 5, and 6 of their report.

Here we will limit our consideration to the Basic level. The group means and standard deviations of the Basic level ratings reported previously by Hambleton and Jones are reproduced in Table 4. Also shown in Table 4 are the NAEP scale scores implied by the group means. A graphical display of the Basic achievement level scale score values by grade and group of raters is provided in Figure 9. The scale score values are based on the empirical test characteristic curves for the composites at grades 4, 8, and 12 that were prepared by ETS and provided to NAGB on November 8, 1990.

In the NAEP scale score metric, the group means have a range of 30 points at grade 4, 54 points at grade 8, and 39 points at grade 12. The Basic achievement levels defined in round 3 by two of the groups of raters at grade 8 are as high or higher than the Basic level defined by any of the four grade 12 groups. The grade 4 Basic achievement level implied by the ratings of the most lenient group (group 1) the 25th percentile on the 1990 math composite, whereas the level implied by the ratings of the most stringent group (group 2), corresponds to the 62nd percentile. Thus, the percentage of students performing below basic would differ by 37% at grade 4 as a function of the group of raters. At the other two grades, the corresponding differences in the percentage of students scoring below Basic as a function of the group of raters is even higher (45% at grade 8 and 54% at grade 12).

On previous NAEP scales the standard deviation for the combined sample of 4th, 8th, and 12th grade students has been set at 50. The within-grade standard deviations are considerably smaller, typically less than 40 and sometimes around 30. These

standard deviations provide an additional perspective for judging the group-to-group variability in Basic achievement levels shown in Figure 9. Group-specific Basic achievement levels differ by as much as roughly a standard deviation or little less at grades 4 and 12, and by substantially more than a standard deviation at grade 8.

The results in Figure 9 suggest that achievement levels are likely to be subject to wide swings from one panel of judges to another. This finding is consistent with previous research on standard setting that shows that standards vary greatly as a function of the groups that provide the judgments as well as the specific rating procedures used. In our judgment, **the large group-to-group variability is another indication that the achievement levels are not sufficiently robust to be credible.**

CRITERION-REFERENCED VS. NORMATIVE ACHIEVEMENT LEVELS

In NAGB's policy framework for setting achievement levels, a distinction was made between the anchor points of 200, 250, 300, etc., that have been used in the past, and the achievement levels to be established as goals. As was indicated, the anchor points were "derived from the distribution of test results themselves, not from any prior judgment of what students ought to know" (NAGB, May 10, 1990, p. 4). That is, the anchor points are basically normative achievement levels for the population of students at a given point in time. In contrast, the achievement levels were intended to define "what performance ought to be" (NAGB, May 10, 1990, p. 5). In the jargon of educational measurement, the achievement levels are intended to be criterion-referenced rather than norm-referenced.

1. Relationship of Achievement Levels to Normative Performance.

The distinction between norm- and criterion-referenced performance standards led us to question the degree to which the achievement levels are related to normative student performance, and to the extent the levels set for items are related to other factors, such as the content or process categorization of items. In an effort to investigate the relationship of achievement level item ratings and normative student performance, the round 4 ratings adjusted for the average change from round 4 to round 5 ratings (see, Hambleton and Jones, December 7, 1990) were plotted against the item p-values. Figures 10, 11, and 12 present the scatterplots of grade 4 adjusted round 4 Basic, Proficient, and Advanced achievement levels with the p-values. Also shown at the bottom of each of these figures is the number of items, the correlation between the adjusted round 4 achievement level ratings and the item p-values, and the correlation of the logit transformations of those two variables.

An inspection of Figures 10, 11, and 12 shows that there is a strong correlation between the item p-values and each of the achievement level ratings at grade 4. The relationship is reasonably linear for the Basic and Proficient ratings but less so for the Advanced ratings due to a marked ceiling effect with the latter ratings. The correlation following logit transformations of the Advanced level proportion and the p-values (.80) is similar in magnitude to the correlations of the Basic and Proficient level ratings with the p-values.

Analogous scatterplots are shown in Figures 13, 14, and 15 for grade 8 and in Figures 16, 17, and 18 for grade 12. For ease of comparison, the correlations for all three grades and all three achievement levels are summarized in Table 5. The relationships between the achievement level ratings of items and the item p-values are noticeably stronger at the two higher grades than they are at grade 4.

With correlations as high as .9 it seems clear that judges are sensitive to the difficulty of items for students. It is unknown, however, whether this sensitivity is due to their awareness of the item p-values and item test regressions, to the raters' informal and implicit understanding of difficulty, or to substantive judgments of the items. Similar analyses of round 1 ratings which were obtained before judges were given normative data on student performance would be revealing in this regard. **Nonetheless, the strength of the relationships suggest that the achievement level ratings are substantially influenced by normative considerations.**

2. Predictions of Basic Level Ratings from Item P-Values.

The regressions of the adjusted round 4 Basic level ratings on the item p-values are shown in Table 6 along with the means and standard deviations of the p-values at each grade. Regressions for Proficient and Advanced ratings were also computed. In addition, regressions were computed for all three rating types using logit transformations. Since the Basic level regressions were nearly linear, we focus only on those simple regressions here.

The slope of the regression is substantially higher at grade 12 than at the other two grade levels. This steeper slope is primarily due to a larger standard deviation of the Basic ratings at grade 12 than at the other two grades (24.1 at grade 12 vs. 13.0 and 12.9 at grades 4 and 8, respectively), but the correlation was also highest in grade 12 (.93 at grade 12 vs. .78 and .90 at grades 4 and 8, respectively). All three slopes are less than 100 -- the value that would directly transform the proportions used to express p-values into the percent metric used to express the Basic ratings.

One of the implications of these regression equations is that the Basic ratings of difficult items (low p-values) will tend to be higher at grades 4 and 8 than at grade 12. The converse is true for items that are extremely easy (say, $p > .8$). This difference in predictions is illustrated for selected p-values in Table 7.

As was noted above, the contrast in predictions for grades 8 and 12 helps explain the apparent anomaly that items common to both grades tend to have higher achievement level ratings at grade 8 than at grade 12. The items unique to grade 12 tend to be more difficult than the items common to both grades, whereas the converse is true at grade 8. As can be seen in Table 7, an item with a p-value of .2 and .4 at grades 8 and 12, respectively, would be expected to be have a slightly higher Basic level rating at the lower grade (47.7 vs. 45.1).

The effect of the difference in regression slopes on the grade 8 and 12 ratings of common items is illustrated graphically in Figure 19. On average, the proportion of grade 8 students that correctly answered the 19 items in the reduced item set, that were common to grades 8 and 12, was .37. From the grade 8 regression line shown in Figure 19, it can be seen that the expected Basic achievement level for a grade 8 item with a p-value of .37 is 56.3. The actual average grade 8 achievement level set for the items common to grades 8 and 12 (59) is shown in parentheses and is reasonably close to the value expected from the regression line.

At grade 12 the average p-value for the same 19 items common only to grades was .51. As can be seen in Figure 19, the expected grade 12 Basic rating for those items is 54.7, which is nearly equal to the value of 54 that was actually obtained from the round 4 ratings.

Both grade 8 and grade 12 judges provided Basic ratings that were highly predictable from student performance on the items. As is reflected by the flatter slope, however, the grade 8 judges spread their ratings over a much narrower, and generally higher, range than was used by the grade 12 judges. **We can think of no educational rationale for this difference between the two panels of judges, and hence conclude that it is due to differences between the groups of judges or in the group dynamics during the rating process rather than to legitimate differences in performance that 8th and 12 grade students "ought" to achieve.**

3. Comparisons Among Ratings Across Content and Process Areas.

The question of whether the achievement levels are criterion-referenced is also illuminated by comparing the relationships between the levels and normative performance across the five a priori content and three process classifications. If the

achievement levels are truly criterion referenced, one would expect that they would call for more improvement in some areas than in others, and the relationships described above would therefore differ from one area to another. A finding that those relationships are invariant would imply either that the achievement levels are implicitly normative or that students' current performance falls short of the criterion by equal amounts in every process and content area--an implausible assumption.

Examination of variations across content and process areas yielded mixed results. However, the relationships are so similar in so many instances, particularly in grade 8, that they call into doubt the criterion-referenced interpretation of the achievement levels.

Grade 12 Results for Content Areas. In grade 12, the relationships between basic ratings on normative p-values show moderate variation across content areas in some instances but little or none in many cases.

Simple mean differences between basic ratings and p-values vary modestly across content areas. At one extreme, the average difference in algebra was -.04, indicating that judges expected students at the basic level to perform slightly worse than the current average. Measurement and statistics are at the other extreme, with ratings exceeding average p-values by .08 and .09, indicating that judges expect students at the basic level to perform somewhat better than the current average. Numbers and operations and geometry both showed positive but smaller mean differences.

Examination of regressions with content areas show that these simple mean differences reflect differences in both intercept and slope. As a consequence, the differences in expected improvement are larger in some ranges of normative difficulty than the simple mean differences suggest. The differences in slope are also sufficiently large that the rank ordering of expected improvement (across content areas) changes with level of difficulty. In terms of intercepts, algebra was again at one extreme, with an intercept of -.01, implying that judges expect essentially no improvement over current performance at the hard end of the distribution. In terms of intercepts, however, the other extreme is in this case numbers and operations (.260).

The practical implications of these differences can be seen in Table 8, which shows the predicted basic rating in each content area for three levels of normative difficulty ranging from very difficult (p=.20) to very easy (p=.80). At the easy end of the distribution, the content areas vary little; in all five, the basic levels are equal to or trivially higher than the normative p-value. At the difficult end of the distribution, however, the basic ratings vary considerably. In algebra, the basic ratings are slightly lower than current performance. In contrast, in

numbers and operations, judges would expect a doubling of the success rate, and they would expect nearly as much improvement in measurement and statistics.

In terms of basic levels, then, the question of criterion-referencing boils down to whether a reasonable set of a priori criteria would call for equal and near zero improvements at easy levels, large improvements in difficult numbers and operations, and no improvement in difficult algebra items to reach the basic level of proficiency.

As one moves to the higher achievement levels, the differentiation between subject areas generally becomes smaller. This is illustrated in Table 9, which presents comparable results for the advanced ratings. The advanced ratings imply very large and quite similar improvements at the hard end of the distribution in all five content areas; again, less improvement is expected in algebra than in several other areas, but the differences are small. Differences among subject areas are even smaller at the easy end of the distribution, in part because of a ceiling effect. The interaction between content area and difficulty is also apparent, though hardly striking; note that change in expected performance levels across normative difficulty is steeper in statistics, for example, than in numbers and operations.

Grade 8 Results for Content Areas. In general, grade-8 regression results tended to vary little--considerably less than the grade-12 regressions--in terms of both intercepts and slopes. There was slight variation at the basic level and less yet at the two higher levels.

The grade-8 basic ratings for all content areas other than algebra were within a range of .03 at the difficult end of the distribution and at a p-value of .50 (Table 10). Algebra was the exception, although less so than in grade 12. In the case of very hard items, judges expected somewhat less improvement in algebra than in the other areas, although in grade 8 they did expect substantial improvement. At the easy end of the distribution, algebra no longer stands out, because of the slightly greater variation in ratings across content areas.

Grade-8 proficient ratings show even less variation across content areas (Table 11). At all levels of normative difficulty, the judges ratings were within a range of .04 across all content areas, including algebra.

Results for Process Areas. The mathematics items used in the ratings were classified into three process categories as well: concepts, procedures, and problem solving. As in the case of content areas, one might expect truly criterion-related achievement levels set different standards for improvement in the three content areas. Such differences would again be apparent in

variations in the regressions of the achievement levels on normative p-values across the three process areas.

In both grades 8 and 12, the relationships between the achievement levels and normative p-values showed almost no variation across process areas. For example, in grade 8, judges expected fairly large and similar improvements at the hard end of the distribution in all three process areas, but no improvement in any at the easy end of the distribution (Table 12). Even less variation was apparent in the case of grade 12 advanced ratings, and the implied p-values are essentially identical across the three process areas (Table 13).

PRECISION OF MEASUREMENT AT DESIGNATED ACHIEVEMENT LEVELS

For achievement levels to provide useful information about student performance in comparison to fixed standards, it is essential that the assessment instrument be designed to provide reasonable measurement precision at those critical levels. If the achievement levels were set in advance of instrument design as a part of the item developmental process, steps would normally be taken to assure that there were adequate numbers of items that function most effectively at student performance levels corresponding to the designated achievement levels. Such an approach obviously was not possible in the present situation because the NAGB achievement level work started with an existing pool of items. It is nonetheless important to assess the adequacy of the item pools for providing the desired precision in measurement at the achievement levels that were set.

The standard error of measurement provides an indication of the precision or dependability of an assessment instrument. Where item response theory is used to scale an assessment, as it is in the case of NAEP, the standard error of measurement can be computed for particular points on the scale. We used the item parameters for the items in the reduced item pool provided by ETS to compute standard error curves for NAEP scale scores, covering the range including the three achievement levels at each grade. Plots of the standard error curves are shown in Figures 20, 21, and 22, for grades 4, 8, and 12 respectively.

It should be emphasized that the standard error curves shown in Figures 20, 21, and 22 apply to the item pools and not to the set of items that would be administered to an individual student. A given student would be expected to respond to only about three sevenths of the items in the reduced item pools because a test booklet contains three of the seven blocks of items in the math assessment. Thus, the test information function for a booklet would be expected to be about three sevenths of that for the reduced pool. The standard errors for scores based on a booklet would be expected to be larger than those shown in Figures 20 through 22 by a factor of approximately 1.5 (which is equal to the square root of seven thirds). It is the shape, or relative height

of the curves that is relevant for present concerns rather than the absolute magnitude of the standard errors.

The letter "A" with an arrow above it in each of the three figures, showing the standard error of measurement curves, shows the approximate level where the Advanced achievement level would be set. As can be seen in Figures 20 and 21, at both grades 4 and 8, the Advanced level is in a score range where the standard errors have begun to increase. This is also true at grade 12, but to a lesser extent. At grade 4 the standard error of measurement is about 1.6 times as large where the Advanced achievement level is set, as it is where the NAEP item pool provides the best measurement. The corresponding figure at grade 8 is about 1.7. That is, the measurement precision is not as good at the grade 4 and grade 8 Advanced levels as it is for lower score levels.

The implication of the results displayed in Figure 20 though 22 is that **NAEP, as presently designed, is not aimed at a high enough level to provide good measurement at the Advanced achievement levels**, particularly at grades 4 and 8. These results provide support for the testimony of Gregory Anrig, President of ETS, for the NAGB Public Hearings on NAEP Achievement Standards on January 8, 1991. As Anrig noted, "it is difficult to find sufficient numbers of exercises in the current assessment ... that embody what we mean by 'advanced' performance. In effect the standard outstrips the current test content" (p. 2 of testimony).

PROCESS PROBLEMS AND ADJUSTMENTS

As is discussed in depth in the evaluation report prepared by Stufflebeam, Jaeger, and Scriven (1991), a number of problems were encountered in the implementation of the standard setting procedures. Among the more salient of these problems was the lack of sufficient time to complete all the planned steps at the Vermont meeting. In addition to having insufficient time for discussion prior to round 3 ratings, several other concerns surfaced during the Vermont meeting. Concerns about confusion regarding the definitions of achievement levels, confusion about how items should be rated that judges considered inappropriate, and the lack of prior pilot testing of procedures are among those discussed by Stufflebeam, et al. Since Stufflebeam, et al. cover these procedural issues in some detail, we will focus our attention on issues related to the conversion of the ratings into achievement levels.

According to Hambleton's December 13, 1990 memorandum to Roy Truby regarding "Recommended Adjustments in the Grades 4, 8, and 12 Achievement Levels," there are a number of possible proportion correct values that might be used to set achievement levels on the NAEP scale. Hambleton suggests that the Round 5 results should be adjusted for the reduced item pool (i.e., the exclusion of higher-order thinking skills and estimation items) and the use of medians rather than means. At grade 4 only, an additional adjustment for changes in the group of raters who provided

ratings for rounds 4 and 5 (i.e., Washington group) in comparison to those who provided ratings only in rounds 1, 2, and 3 (i.e., Vermont only group), was considered.

The use of medians rather than means is desirable given the skew in some of the distribution of ratings. The adjustments for excluded items and possibility of adjustments for differences between Vermont only and Washington raters, on the other hand, while reasonable approaches to solve problems after the fact, leave important questions unanswered.

The decision to exclude items was based, at least in part, on the fact that the higher-order thinking skills and estimation items, while part of the national administration, were not included in the booklets used for the trial state assessment at grade 8. The achievement levels were intended for use in reporting results of the trial state assessment as well as the national administration, but it was judged inappropriate for the trial state assessment results to use standards set on an item pool that included items that were not administered in that assessment. Hence, it was recommended that the higher-order thinking skills and estimation items be excluded when the ratings were used to determine the achievement levels on the NAEP scale.

It is important to note in this regard that the excluded items were part of all four rounds of ratings of individual items. Moreover, judges were not told that those items would be excluded when they made their global ratings in round 5. The numbers of excluded items were 34, 54, and 59 at grades 4, 8, and 12, respectively. These numbers correspond to almost a third of the rated items at grade 4 and roughly two-fifths of the rated items at grades 8 and 12 (31%, 39%, and 41% at grades 4, 8, and 12, respectively).

In retrospect, it clearly would have been better to rate only the items in the reduced item set if higher-order thinking skills and estimation items were to be excluded in the final determination of the achievement levels. After the fact adjustments of round 5 ratings based on the difference in average ratings of included and excluded items at round 4, while intuitively reasonable, are no substitute for ratings that are explicitly tied to the set of items to be used.

It should also be noted that the excluded items are ones that content experts judge to be of considerable importance. They are consistent with directions that mathematics educators are urging that the assessment move for the future. Although these views are open to debate, **excluding the higher-order thinking skills and estimation items may establish achievement levels on content that is, in the judgment of some, more relevant to the past than the future.** In addition, the exclusion of those items exacerbates the previously noted problem of obtaining adequate measurement at the Advanced achievement level because

some of the excluded items contribute most to the assessment of high performing students.

The adjusted round 5 percent correct values for the three achievement levels are listed in Table 14 by grade. Two sets of values are shown for grade 4. These correspond to the values with and without the adjustment for differences between Vermont only and Washington raters. Also shown in Table 14 are two sets of scaled scores that correspond to the percent correct values. The first set of scaled score values is based on the empirical test characteristic curves for the composite provided to NAGB by ETS in November and include the higher-order thinking skills and estimation items.

The second set of scale score values is based on the test characteristic curves that were computed using the item parameters for the items in the reduced item pools at each grade. As can be seen, the achievement levels defined in scale score units are fairly similar for the reduced item pool test characteristic curves and the ones based on the complete sets of rated items.

The potential adjustment for differences between the Vermont only and Washington raters at grade 4 makes a somewhat greater difference. Judging from experience with earlier NAEP scales, an increase as large as 7 points, which occurs at the Basic achievement level with the reduced item pool test characteristic curve, is likely to correspond to roughly a fourth of a within-grade standard deviation. A difference as large as a quarter of a standard deviation obviously could change the percent of students scoring below the basic achievement level by a substantial amount. As was previously noted in the discussion of Figures 6 and 8, we believe that the differences between achievement levels defined by Vermont only and by Washington raters are too large to ignore. Since we also agree with Hambleton's (December 18, 1990) conclusion that it would be hard to defend the adjustment, it would appear that we are left with no adequately defensible achievement levels for grade 4 students.

It should be recalled that grade 8 raters generally set higher levels than did grade 12 raters for items that were common to both grade levels. If only items common to both grades were used, the grade 8 levels would be higher than their grade 12 counterparts. As can be seen in Table 14, however, the scale scores for each achievement level increase with grade level. As was explained above, the fact that items that were used only at grade 12 tend to be more difficult than those used at both grade 8 and 12, while the converse is true for items that are used at grade 8 but not grade 12, accounts for the grade 12 levels being higher than the grade 8 levels.

Thus, the achievement levels are "coherent" in the sense that Basic (or Proficient, or Advanced) achievement is defined to be

higher at grade 12 than 8 and higher at grade 8 than 4. This "coherent" outcome does not, in our judgment, overcome the "incoherence" of the ratings of items that are common to both grades 8 and 12. There is ample evidence that students perform better on those commonly rated items at grade 12 than at grade 8. As was previously noted, the explanation for a reversal in the ratings is more likely to be found in differences between the panels of raters than in educationally sound reasons for expecting higher performance at grade 8 than 12 for those items.

INTERPRETATIONS OF ACHIEVEMENT LEVELS: HOW VALID AND HOW CREDIBLE?

The previous sections have dealt with specific aspects of the achievement levels, sometimes at a rather detailed and technical level. In this final section, we have attempted to step back and consider the bigger questions. Given what has been learned from analyses reported by Hambleton, the results of the Stufflebeam, Jaeger and Scriven evaluation, our own analyses, and previous discussions of NAEP, how valid and how credible are the achievement levels? Will introducing the achievement levels for the 1990 assessment accomplish the important ends that NAGB identified when the project was launched?

Gordon Ambach, Executive Director of The Council of Chief State School Officers, framed the issue well in his statement presented at the January 8, 1991 NAGB Hearing on the Achievement Levels. He acknowledged the importance of the task and noted that it is critical that the establishment of the levels be credible if they are to serve as National Goals for Education. He also noted that the effort was undertaken subject to some serious constraints, notably the need to work with an assessment that was not designed with the achievement levels in mind and the need to work on an extremely short time-line. With that background, Ambach identified two critical questions that must be answered.

> "One is whether the levels are appropriate and useful for guiding and monitoring mathematics education: Are the statements of criteria for the levels appropriate statements about desirable levels of achievement in math at various grade levels? The second question is whether the process used to develop the levels is credible in terms of assessment development techniques and psychometric practices; can the Board use the levels confidently to represent student achievement in mathematics? (Ambach statement, January 8, 1991, p. 1).

These two questions are closely related because the achievement levels can only be useful if they are credible. Our analyses, as well as those conducted by Stufflebeam, Jaeger, and Scriven (1991), have identified **many grounds for questioning the credibility of the results** -- e.g., reversals between 8th and 12th grade ratings of common items, the removal of 31 to 41% of the rated items after the ratings were completed and without

the knowledge of the raters when they provided their final global ratings, and the attrition of raters between rounds 3 and 4 leaving only a small and non-representative number of judges at grades 4 and 12 for the ratings that counted. These and other results are also relevant to the question of validity.

The Proficient achievement level is intended to "reflect a consensus that students reaching such a level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. At Grade 12 the proficient level will encompass a body of subject-matter knowledge and analytical skills, of cultural literacy and insight, that all high school graduates should have for democratic citizenship, responsible adulthood, and productive work."

Although we have not yet been provided with the official scale scores corresponding to each level or with the percentage of students below each achievement level, it is clear from our analyses that the percentage of students at each grade who performed at the Proficient level or higher in 1990 will prove to be quite low. Based on our approximations to the scale scores corresponding to the round 5 ratings, less than 1 student in 10 at grades 4 and 8 performed at the Proficient level or higher in 1990. Indeed, a majority of students at grade 8 surely performed below the Basic achievement level in 1990 if the round 5 adjusted levels reported by Hambleton are used.

It might be argued that such high achievement levels are precisely what is needed to affect needed changes in education. However, no evidence has been presented that students who perform below the Proficient level, for example, are not well prepared for the next level of schooling. Nor has any evidence been presented that students who score below the Advanced level at grade 12 (which probably includes better than 19 of 20 of the 1990 seniors) are not ready "for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement." **There simply is insufficient evidence to support the validity of the current achievement levels.**

Although the findings reported here are not intended to test the claim that the Proficient levels represent adequate preparation for subsequent schooling, they do offer reasons to doubt its reasonableness. Unless one believes that students need no more preparation to succeed as freshmen in college than in grade 9, the proficient levels should be substantially higher at grade 12 than at grade 8. As noted, if the Proficient levels were based on items rated at both grades, higher levels of performance would be required at grade 8 than at grade 12.

Given the problems that have been identified, we believe that the use of the current achievement levels for reporting national or state-level NAEP results would be a serious error. It would lead policymakers to make insupportable inferences about student performance, and this in turn could lead to undesirable and

possibly counterproductive changes in policy. Moreover, using achievement levels that suffer from such severe flaws could undermine the credibility, not only of the achievement levels, but of NAEP itself. The public and policymakers, in turning to NAEP for invaluable information about student performance, may fail to distinguish between the flaws in the achievement levels and the strength of NAEP itself. Hence, **we recommend that the achievement levels developed so far not be used in any public reporting of national or state-level NAEP results.**

To put these pessimistic conclusions in perspective, we note that the task that NAGB undertook was immense and in many respects unprecedented. We know of no previous effort to establish achievement levels that would serve as "appropriate achievement goals" for the nation as a whole. Moreover, the effort had to be undertaken on a extremely tight time schedule using existing instruments rather than ones designed with the achievement levels in mind. We commend NAGB and the many dedicated professionals who participated in this difficult undertaking. While this first effort did not yield achievement levels that can be used for reporting NAEP results, it did provide a great deal of information that should prove invaluable in future efforts of this sort.

There are a number of fundamental issues worthy of careful consideration if new efforts to set achievement levels are undertaken for 1992 or beyond. The definition of the population of judges, possible stratifications of judges, and sampling from those strata will require detailed consideration. The basic approach also deserves further consideration. The approach of rating individual items in the existing item pool rests on a questionable assumption that complex determinations of student proficiency can be inferred from these ratings. It may be desirable to consider alternative approaches that do not rely exclusively on the aggregation of ratings of individual items. Whatever method is undertaken next, it should provide evidence of robustness, coherence, and the criterion basis of the achievement levels.

Table 1

Average Item Proportion Correct and Round 3 Achievement Level Ratings on the Items Common to Grades 8 and 12.

Number of Items	Student Proportion Correct Grade 8	Student Proportion Correct Grade 12	Judgmental Data: Basic Grade 8	Basic Grade 12	Proficient Grade 8	Proficient Grade 12	Advanced Grade 8	Advanced Grade 12
32[a]	.62	.76	78	73	91	92	97	98
39[b]	.47	.61	66	57	84	84	93	96

a. Thirty two items common to grades 4, 8, and 12.
b. Thirty nine items common to grades 8 and 12 only.

Table 2

Average Item Proportion Correct and Round 4 Achievement Level Ratings on the Items Common to Grades 8 and 12 in the Reduced Item Sets

Number of Items	Student Proportion Correct Grade 8	Student Proportion Correct Grade 12	Judgmental Data: Basic Grade 8	Basic Grade 12	Proficient Grade 8	Proficient Grade 12	Advanced Grade 8	Advanced Grade 12
21[a]	.68	.80	76	77	90	88	98	93
19[b]	.37	.51	59	54	80	76	92	88

a. Twenty one items common to grades 4, 8, and 12 in reduced item set.
b. Nineteen items common to grades 8 and 12 only in reduced item set.

Table 3

Basic Achievement Levels on the NAEP Scale Corresponding to Ratings Obtained by Round for Raters Present in Vermont Only and those Present in Washington[a]

Grade	Meeting Attendance	Number of Raters	Round 1	2	3	4	5
4	Vermont Only	11	222	218	217	NA	NA
4	Washington	11	204	197	203	213	216
8	Vermont Only	3	298	297	296	NA	NA
8	Washington	19	290	296	292	288	280[b]
12	Vermont Only	10	294	292	295	NA	NA
12	Washington	9	290	292	292	301[c]	298

a. Based on ratings of total item pool and the empirical test characteristic curves provided by ETS on November 8, 1990

b. N = 18

c. N = 8.

Table 4

Achievement Level Group Means and Standard Deviations from Round 3 Ratings and Estimated NAEP Scale Scores Corresponding to Group Means*

Grade	Group	Mean	Standard Deviation	NAEP Scale Score
4	1	40.3	5.6	196
4	2	56.3	14.6	226
4	3	47.8	8.0	212
4	4	43.4	9.4	203
4	Total	47.2	11.4	210
8	1	85.2	4.2	324
8	2	82.3	6.6	316
8	3	58.0	7.3	270
8	4	57.7	4.7	270
8	Total	70.9	14.4	292
12	1	66.3	8.3	316
12	2	49.7	6.2	285
12	3	45.9	12.4	277
12	4	57.2	6.6	309
12	Total	54.2	11.1	294

* Mean and standard deviations in percent correct metric are from Hambleton and Jones (December 7, 1990, Tables 4, 5, and 6). The mean percent figures were transformed to the NAEP scale using the empirical test characteristic curves provided to NAGB by ETS on November 8, 1990.

Table 5

Correlations of Item p-values with Round 4 Achievement Level Ratings (Reduced Items Sets)

Observed p-values and Achievement Level Ratings			
	Achievement Level		
Grade	Basic	Proficient	Advanced
4	.78	.80	.74
8	.90	.90	.87
12	.93	.89	.84
Logit Transformations of p-values and of Achievement Levels			
4	.78	.82	.80
8	.90	.91	.88
12	.90	.91	.89

Table 6

Regressions of Adjusted Round 4 Basic Achievement Level Ratings on Item P-Values

Grade	Number of Items	P-Values		Regression	
		Mean	Standard Deviation	Slope	Intercept
4	109	.473	.225	45.40	28.71
8	137	.520	.229	50.62	37.59
12	144	.530	.256	87.23	10.25

Table 7

Predicted Adjusted Round 4 Basic Achievement Level Ratings for Selected Item P-values

P-Value	Standard Deviations From P-Value Mean			Predicted Basic Rating		
	Grade			Grade		
	4	8	12	4	8	12
.2	-1.07	-1.40	-1.29	37.8	47.7	27.7
.4	-.32	-.52	-.51	47.9	57.8	45.1
.6	.56	.35	.27	56.0	68.0	62.6
.8	1.45	1.27	1.05	65.0	78.1	80.0

Table 8

Grade 12 Basic Ratings at Different Levels of Normative p-Values, by Content Area, Predicted from Regressions within Content Area

	p-value		
	0.20	0.50	0.80
Numbers and Operations	0.40	0.60	0.80
Measurement	0.37	0.60	0.82
Geometry	0.29	0.54	0.80
Statistics	0.36	0.59	0.83
Algebra	0.17	0.45	0.72

Table 9

Grade-12 Advanced Ratings at Different Levels of Normative p-Values, by Content Area, Predicted from Regressions within Content Area

	p-value		
	0.20	0.50	0.80
Numbers and Operations	0.86	0.91	0.96
Measurement	0.86	0.91	0.97
Geometry	0.85	0.91	0.97
Statistics	0.81	0.89	0.98
Algebra	0.81	0.88	0.96

Table 10

Grade-8 Basic Ratings at Different Levels of Normative p-Values, by Content Area, Predicted from Regressions within Content Area

	p-value		
	0.20	0.50	0.80
Numbers and Operations	0.50	0.65	0.80
Measurement	0.49	0.65	0.81
Geometry	0.51	0.62	0.73
Statistics	0.48	0.62	0.76
Algebra	0.42	0.58	0.74

Table 11

Grade-8 Proficient Ratings at Different Levels of Normative p-Values, by Content Area, Predicted from Regressions within Content Area

	p-value		
	0.20	0.50	0.80
Numbers and Operations	0.73	0.82	0.90
Measurement	0.71	0.81	0.90
Geometry	0.74	0.80	0.87
Statistics	0.71	0.80	0.88
Algebra	0.70	0.79	0.88

Table 12

Grade-8 Basic Ratings at Different Levels of Normative p-Values, by Process Area, Predicted from Regressions within Process Area

	p-value		
	0.20	0.50	0.80
Concepts	0.50	0.64	0.78
Procedures	0.49	0.64	0.79
Problem Solving	0.45	0.61	0.76

Table 13

Grade-12 Advanced Ratings at Different Levels of Normative p-Values, by Process Area, Predicted from Regressions of Logits within Process Area

	p-value		
	0.20	0.50	0.80
Concepts	0.85	0.92	0.96
Procedures	0.86	0.92	0.95
Problem Solving	0.84	0.91	0.95

Table 14

Achievement Levels for NAEP 1990 Math Assessment

Grade	Level	Percent Correct	Approximate Scale Score	
			All Rated Items[a]	Reduced Item Pool[b]
4[c]	Basic	50.8	217	219
	Proficient	75.7	257	261
	Advanced	90.6	292	297
4[d]	Basic	54.7	222	226
	Proficient	79.1	263	267
	Advanced	92.0	301	303
8	Basic	60.1	274	276
	Proficient	80.2	311	312
	Advanced	92.0	351	346
12	Basic	52.8	291	293
	Proficient	78.7	339	340
	Advanced	89.6	363	365

a Based on empirical test characteristic curves given to NAGB by ETS on November 8, 1990.

b Based on test characteristic curves computed from item parameters for the reduced item pool.

c Grade 4 ratings without final adjustment for the population of raters (based on Hambleton, Dec. 18, 1990 memorandum to Truby).

d Grade 4 ratings with the adjustment for the population of raters (based on Hambleton, Dec. 18, 1990 memorandum to Truby).

Figure 1

Frequency Distribution of Grade 12 Minus Grade 8 Basic Ratings for 71 Items Common to Both Grades (Round 3)

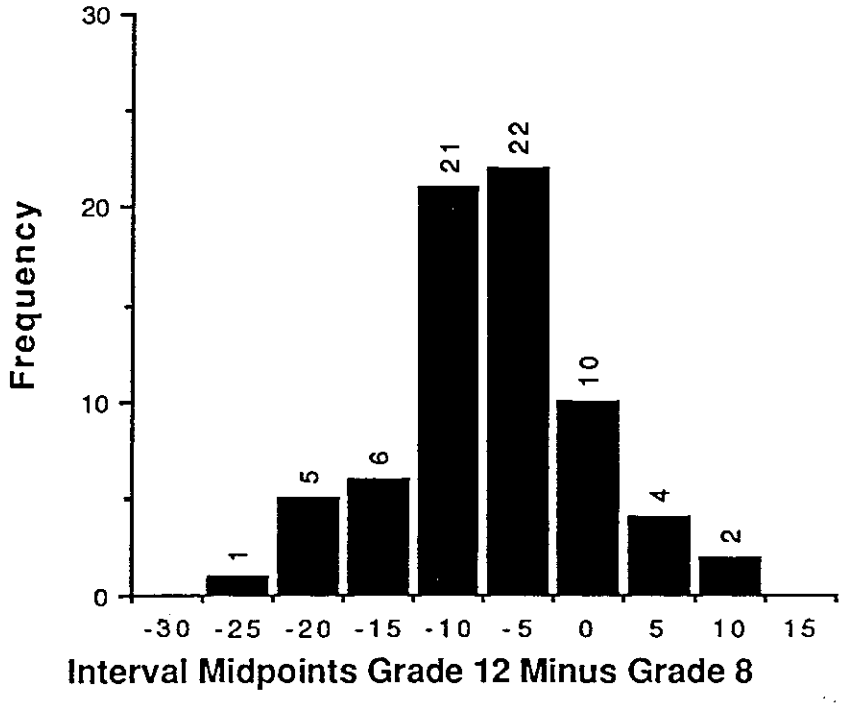


Figure 2

Percent Correct and Round 3 Achievement Levels for 39 Items Common to Grades 8 and 12 Only

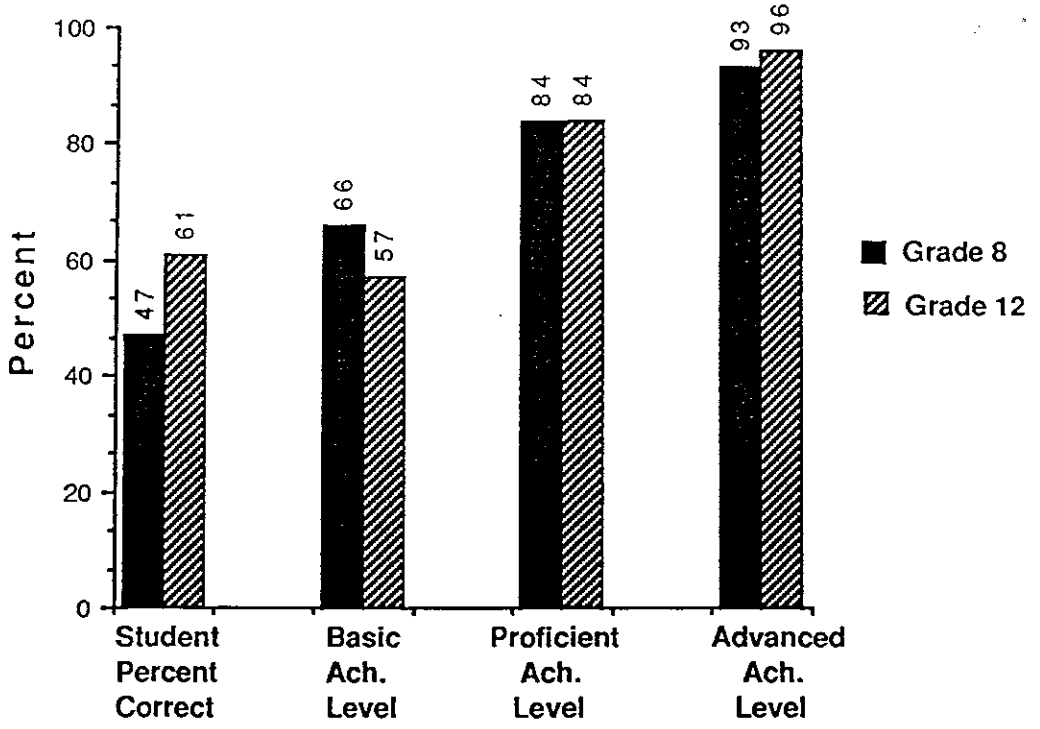


Figure 3

Percent Correct and Round 4 Achievement Levels for 19 Items Common to Grades 8 and 12 Only (Reduced Item Set)

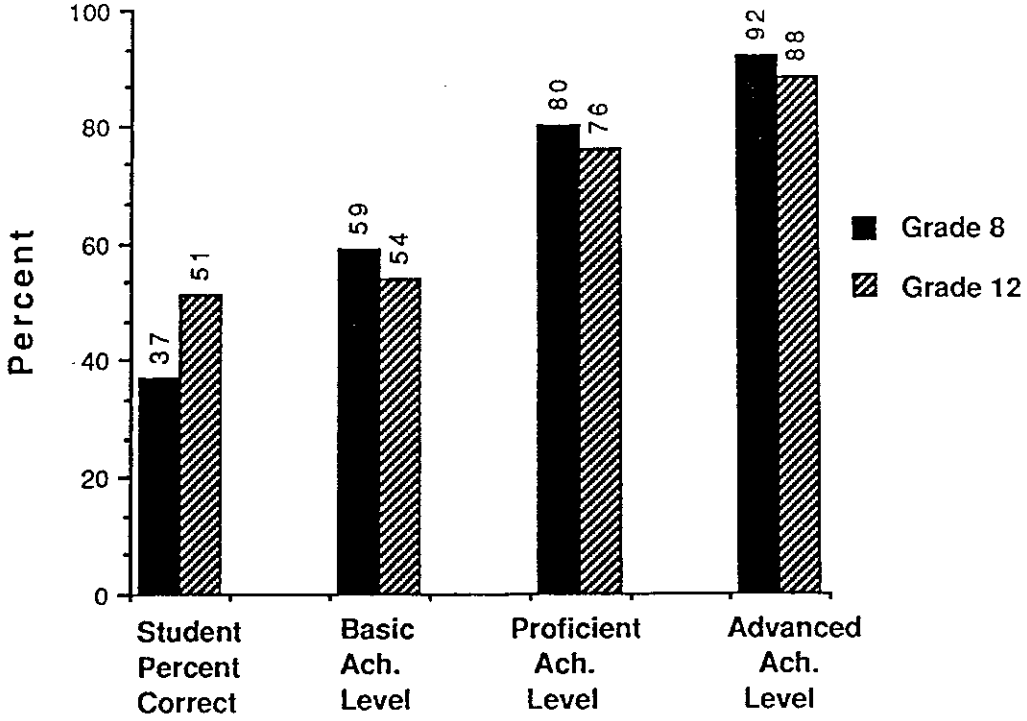


Figure 4

Percent Correct and Round 3 Achievement Levels for 32 Items Common to Grades 4, 8, and 12

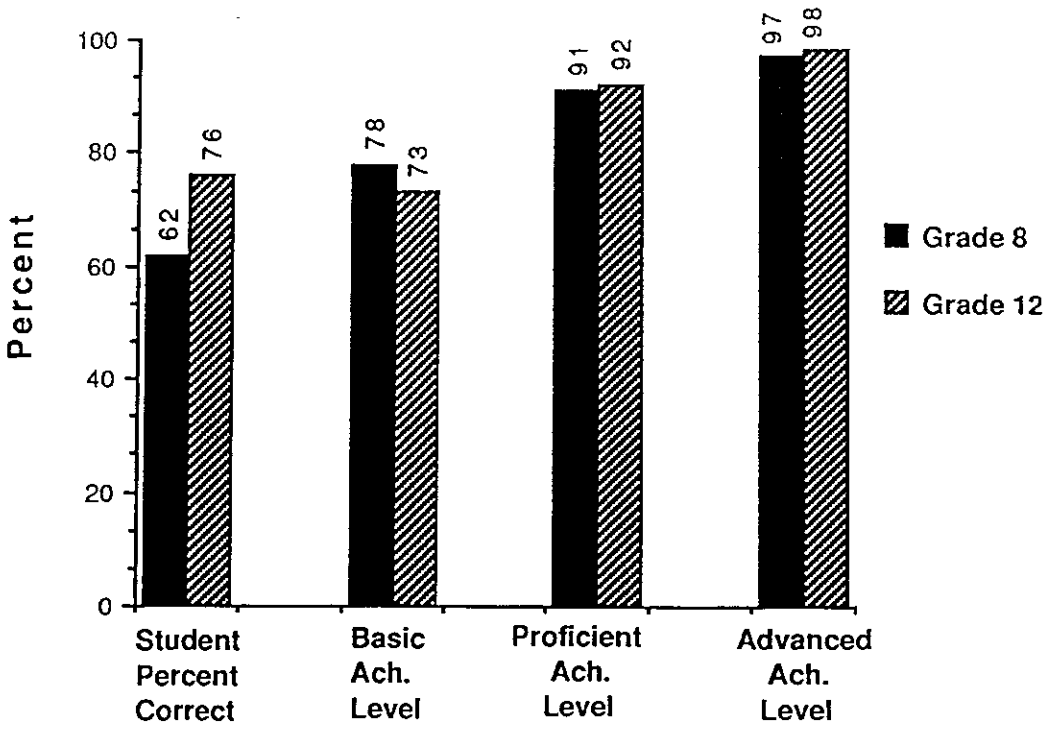


Figure 5

Percent Correct and Round 4 Achievement Levels for 21 Items Common to Grades 4, 8, and 12 (Reduced Item Set)

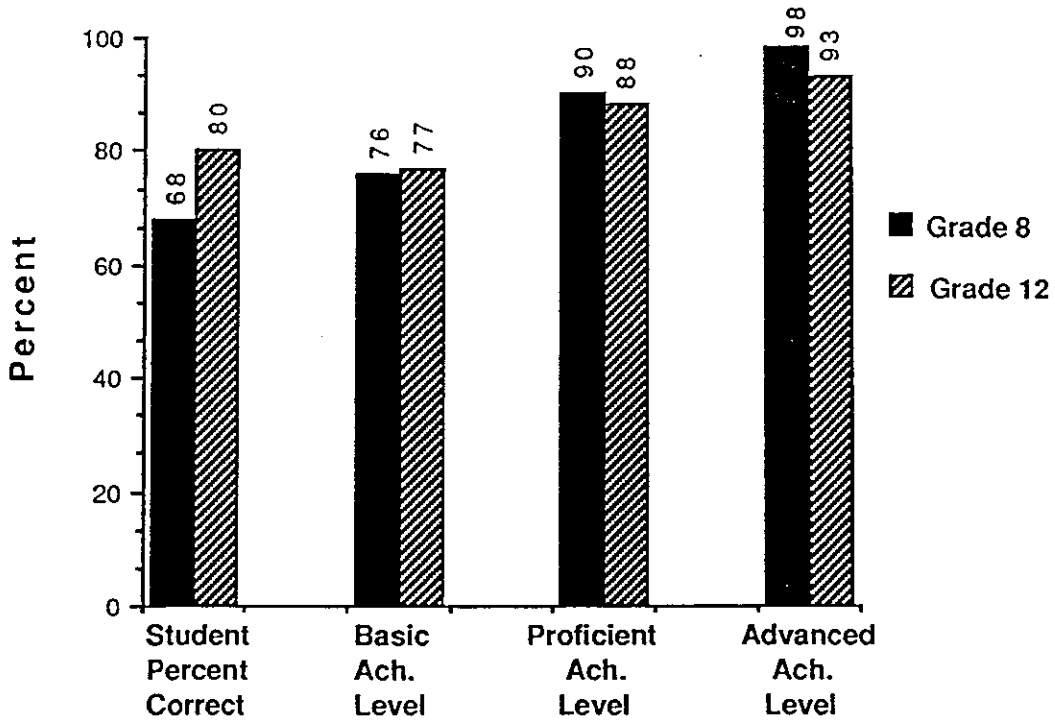


Figure 6

Implied Basic Achievement Levels by Grade, Round, and Meeting Attendance

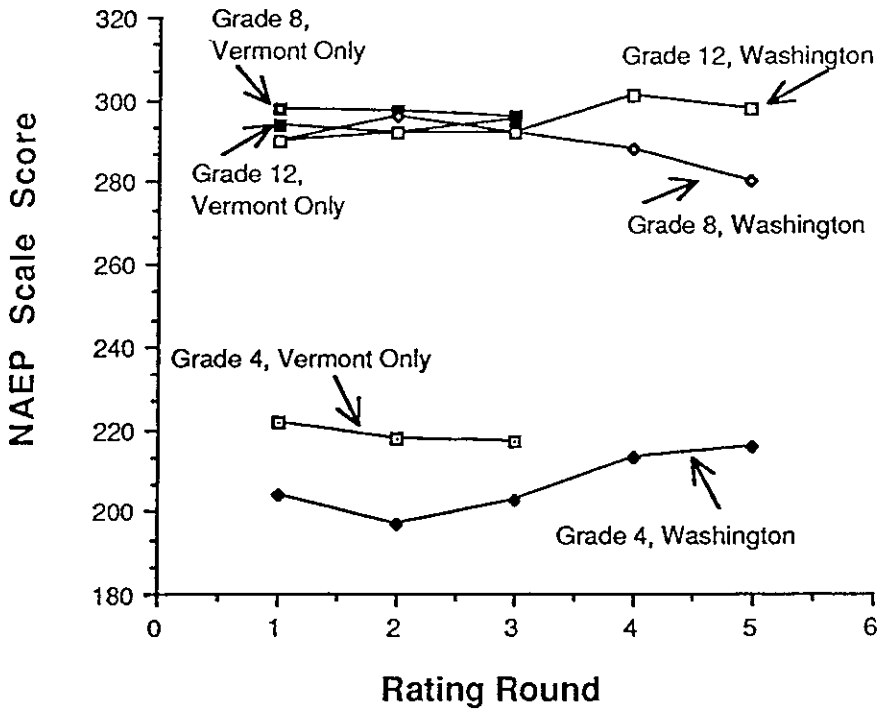


Figure 7

Basic Achievement Levels by Round and Meeting Attendance (Grades 8 and 12)

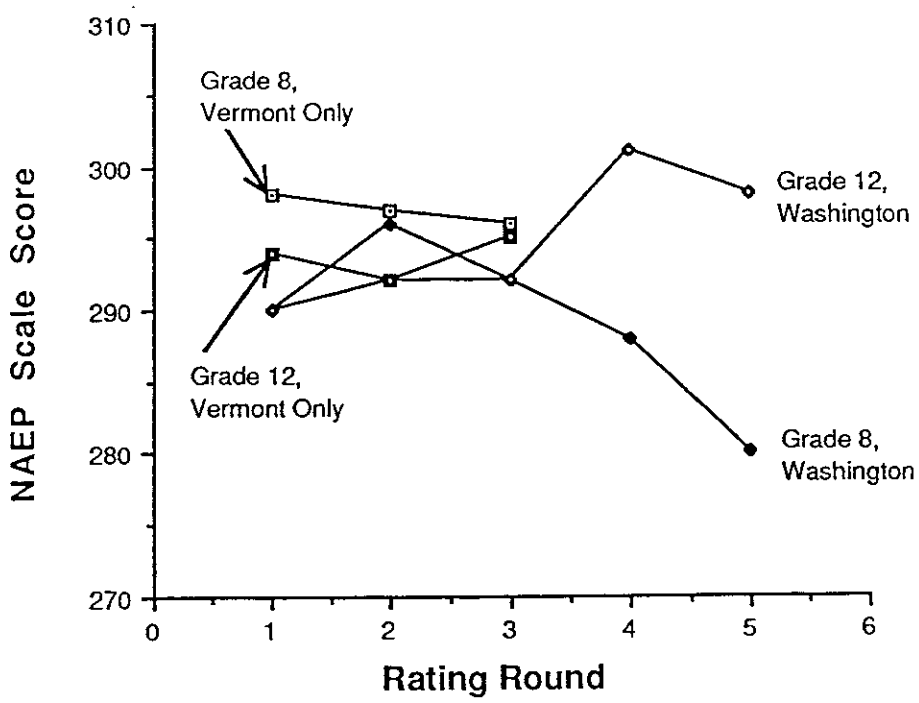


Figure 8

Basic Achievement Levels by Round and Meeting Attendance (Grade 4)

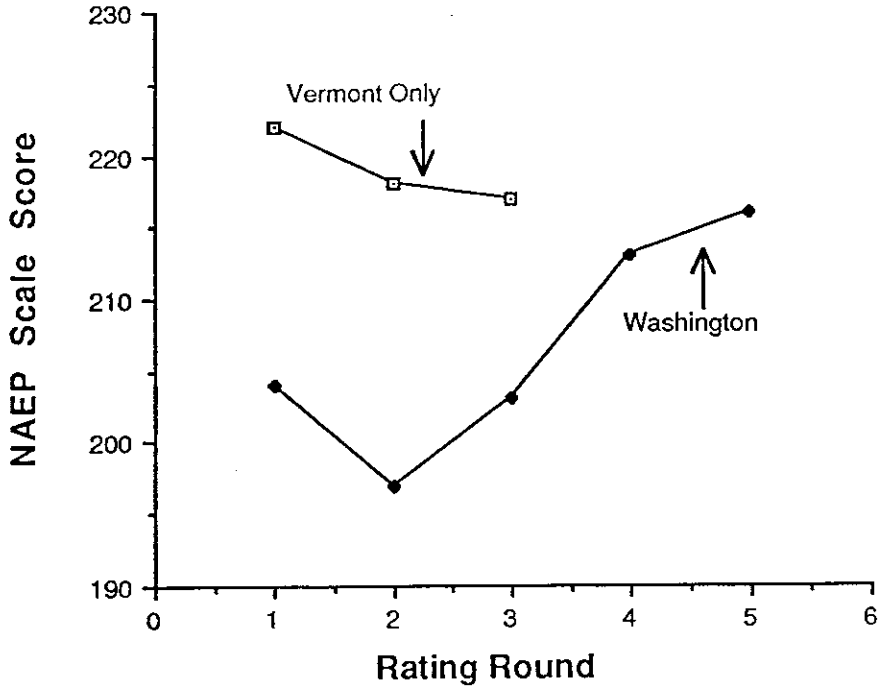


Figure 9

Scale Scores Corresponding to Round 3 Basic Ratings by Grade and Group of Raters

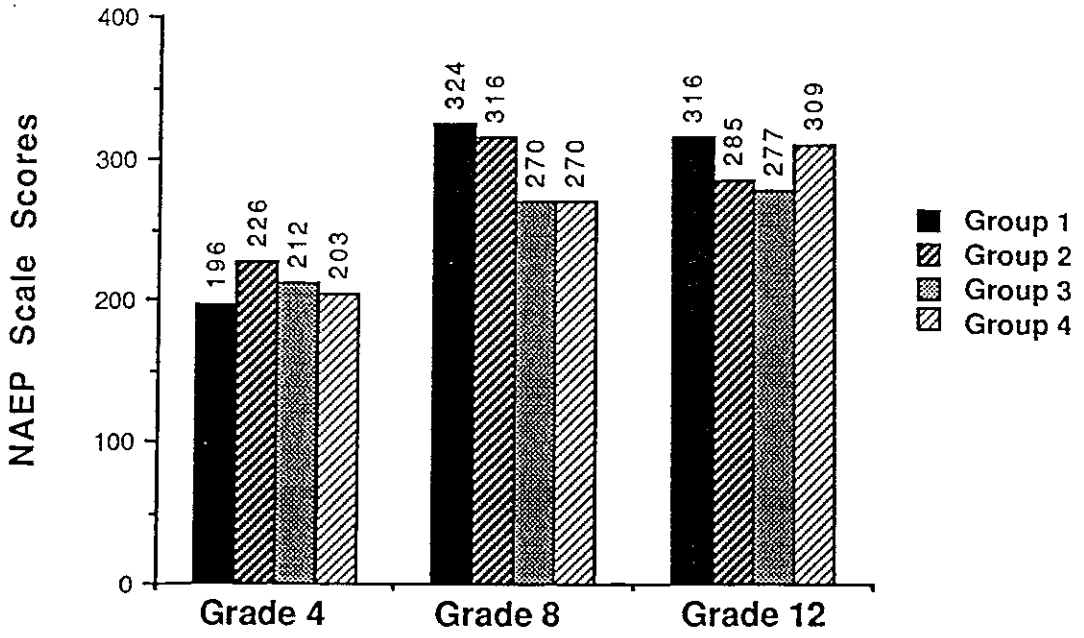
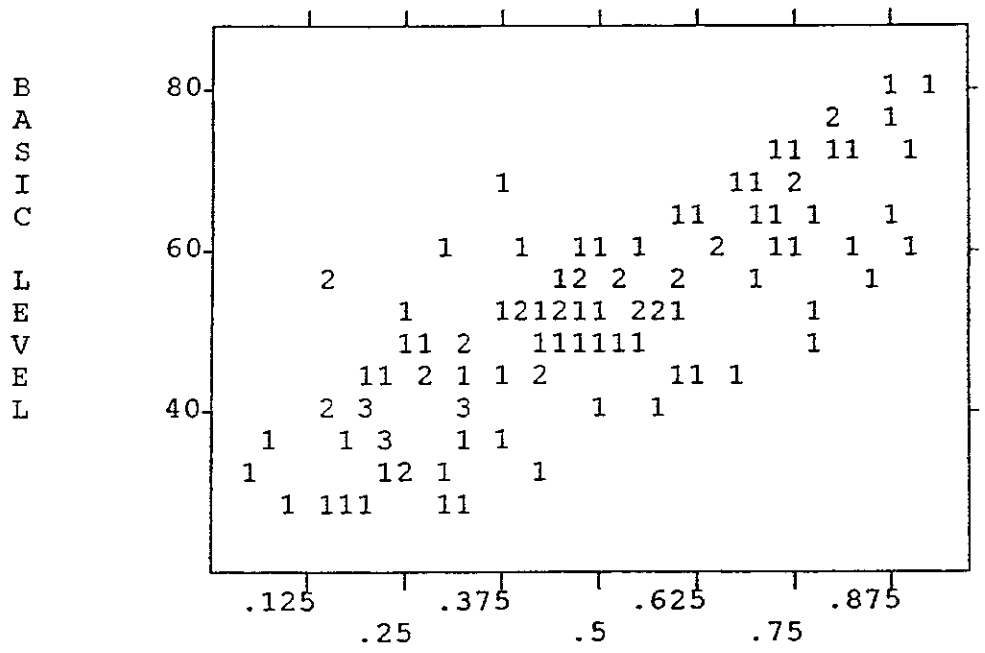


Figure 10

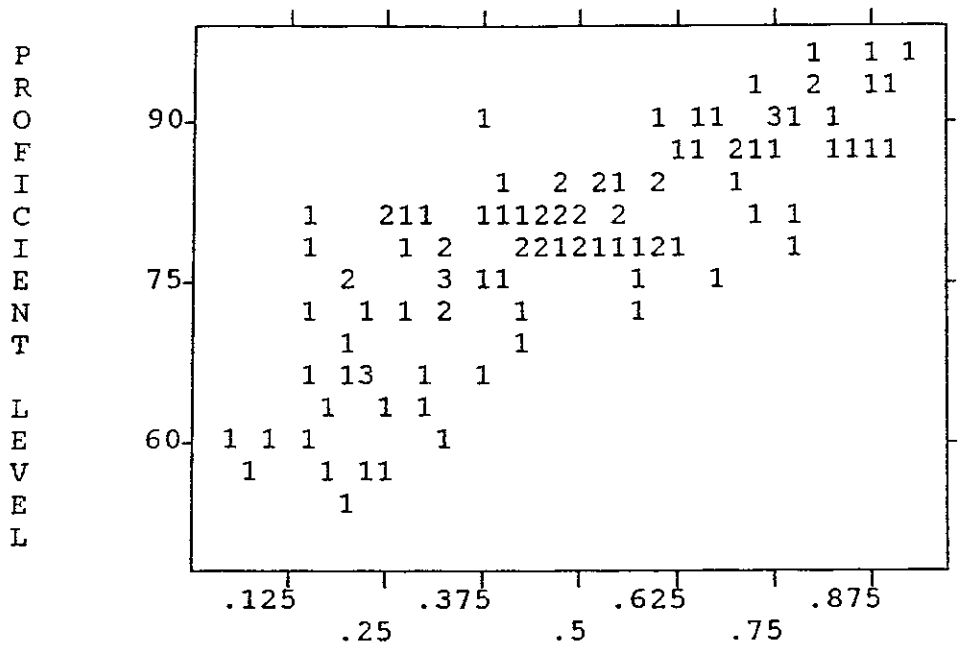
Plot of Adjusted Round 4 Basic Achievement Levels with Item P-Values for Grade 4



Correlation = .78, Number of items = 109. (correlation of logit transformations of two variables = .78)

Figure 11

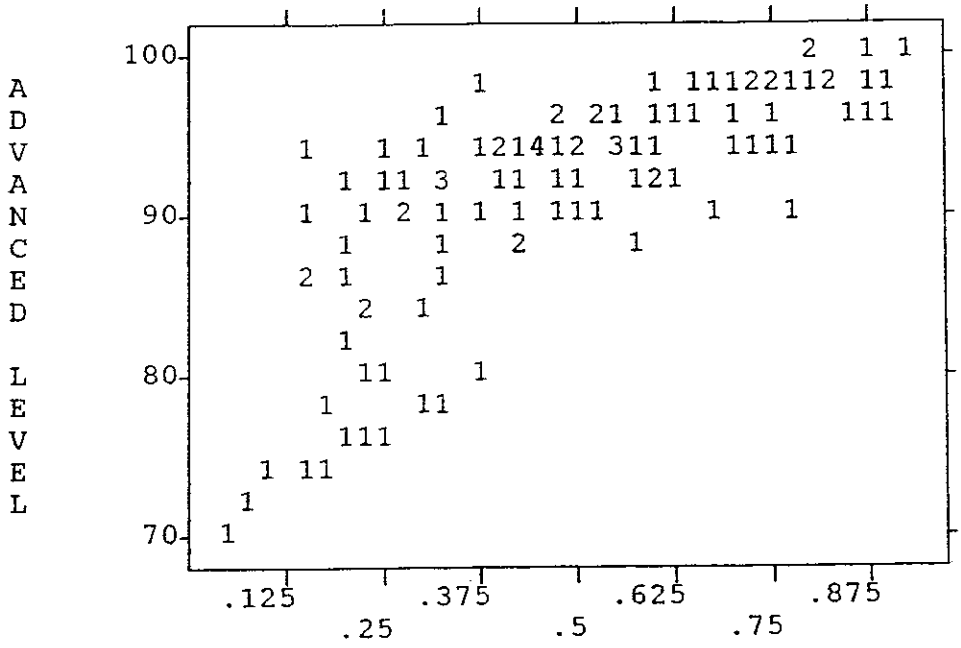
Plot of Adjusted Round 4 Proficient Achievement Levels with Item P-Values for Grade 4



Correlation = .80, Number of items = 109. (correlation of logit transformations of two variables = .82)

Figure 12

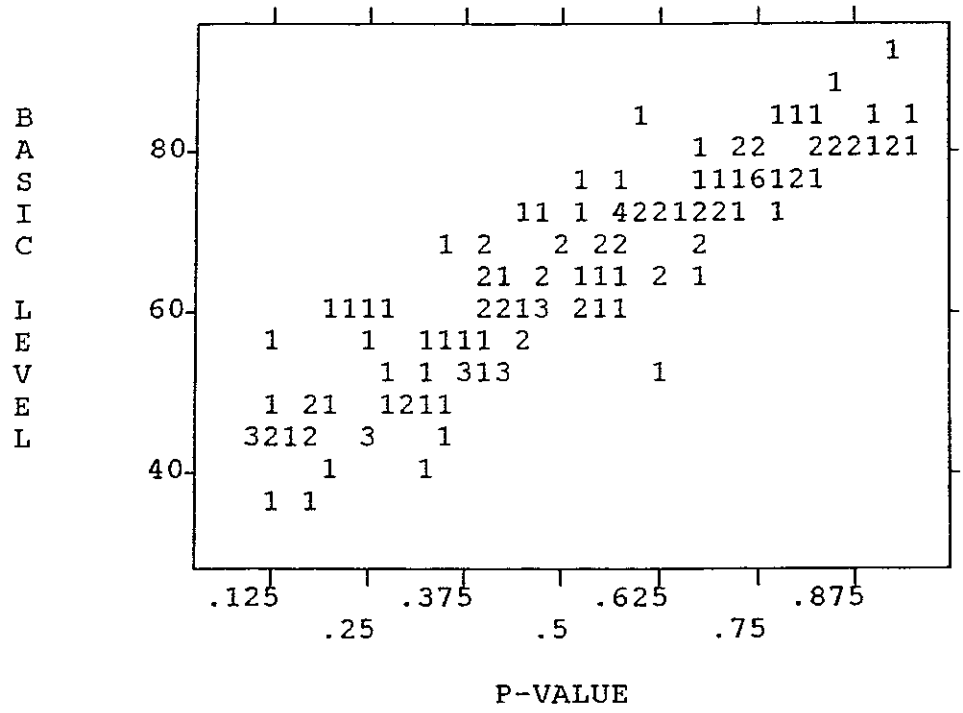
Plot of Adjusted Round 4 Advanced Achievement Levels with Item P-Values for Grade 4



Correlation = .74, Number of items = 109. (correlation of logit transformations of two variables = .80)

Figure 13

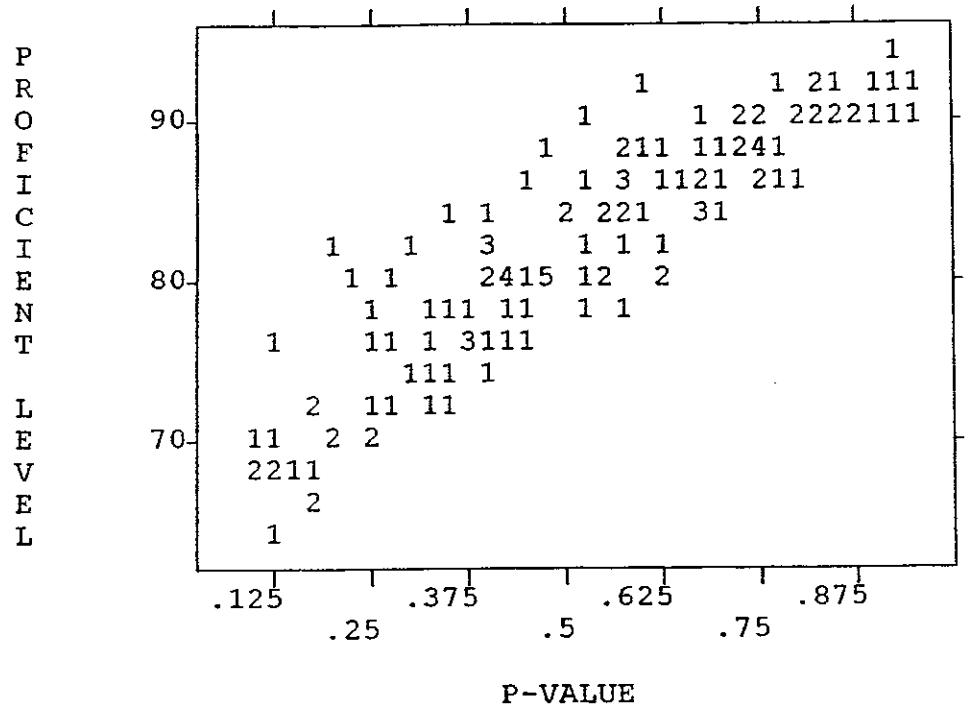
Plot of Adjusted Round 4 Basic Achievement Levels with Item P-Values for Grade 8



Correlation = .90, Number of items = 137. (correlation of logit transformations of two variables = .90)

Figure 14

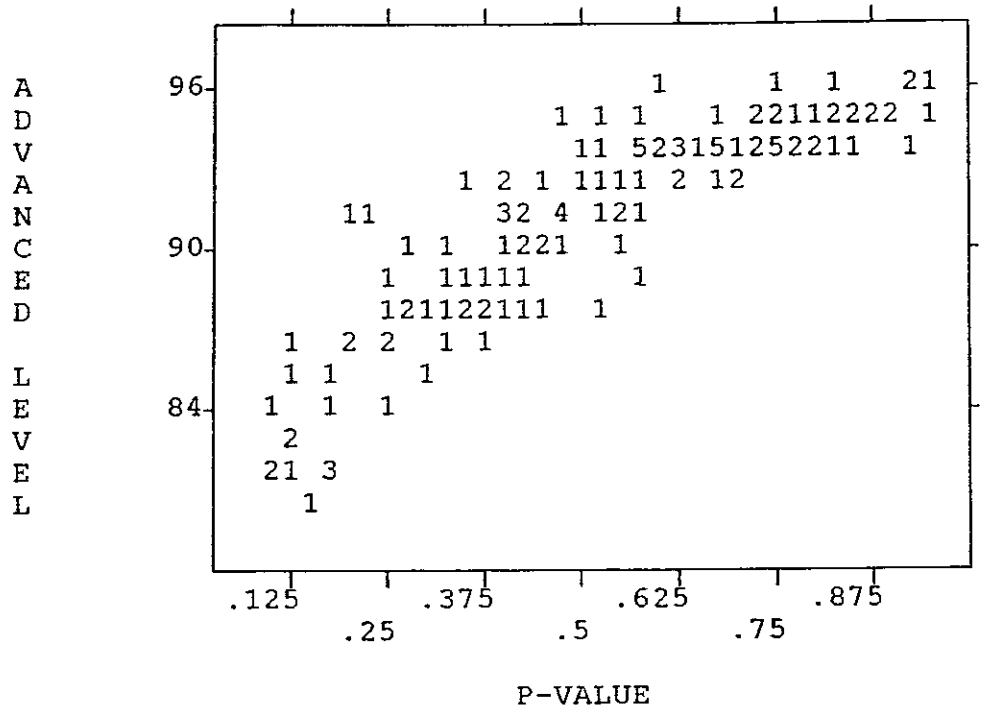
Plot of Adjusted Round 4 Proficient Achievement Levels with Item P-Values for Grade 8



Correlation = .90, Number of items = 137. (correlation of logit transformations of two variables = .91)

Figure 15

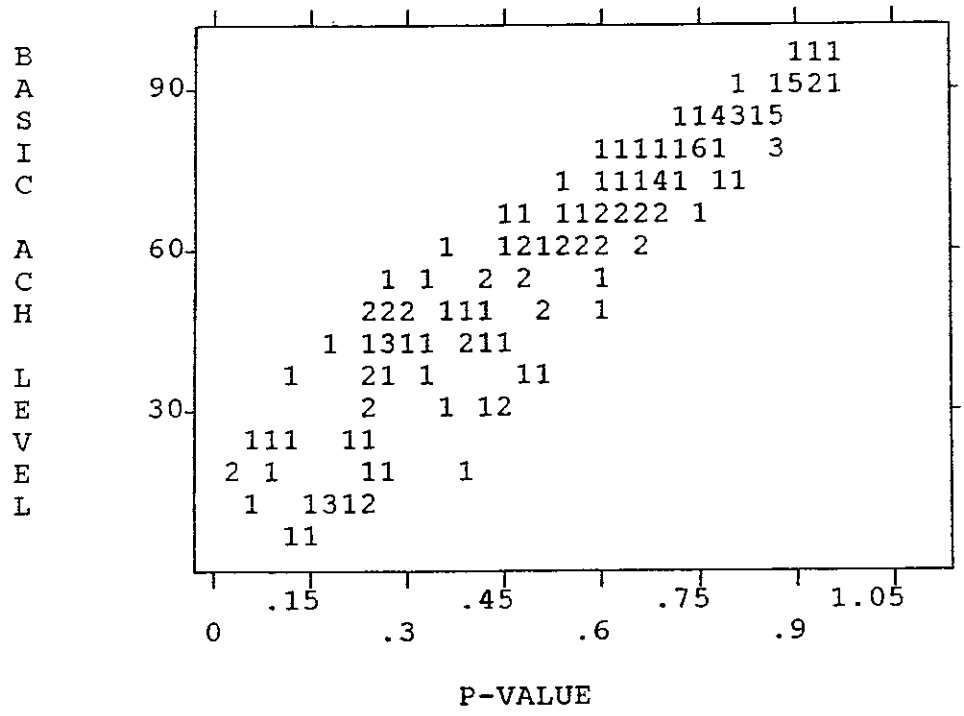
Plot of Adjusted Round 4 Advanced Achievement Levels with Item P-Values for Grade 8



Correlation = .87, Number of items = 137. (correlation of logit transformations of two variables = .88)

Figure 16

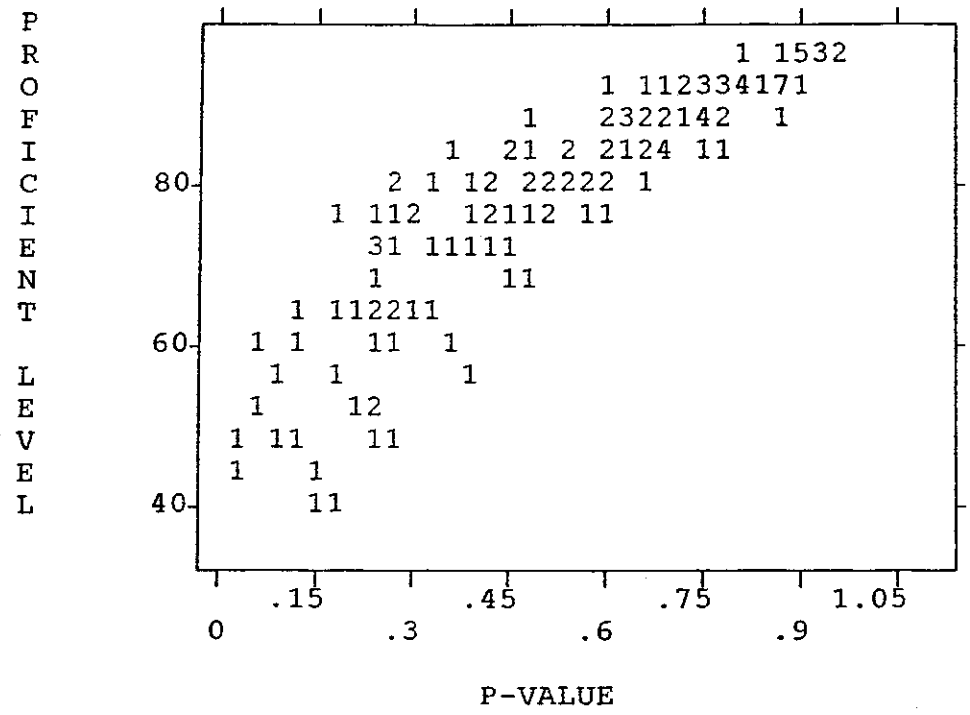
Plot of Adjusted Round 4 Basic Achievement Levels with Item P-Values for Grade 12



Correlation = .93, Number of items = 144. (correlation of logit transformations of two variables = .90)

Figure 17

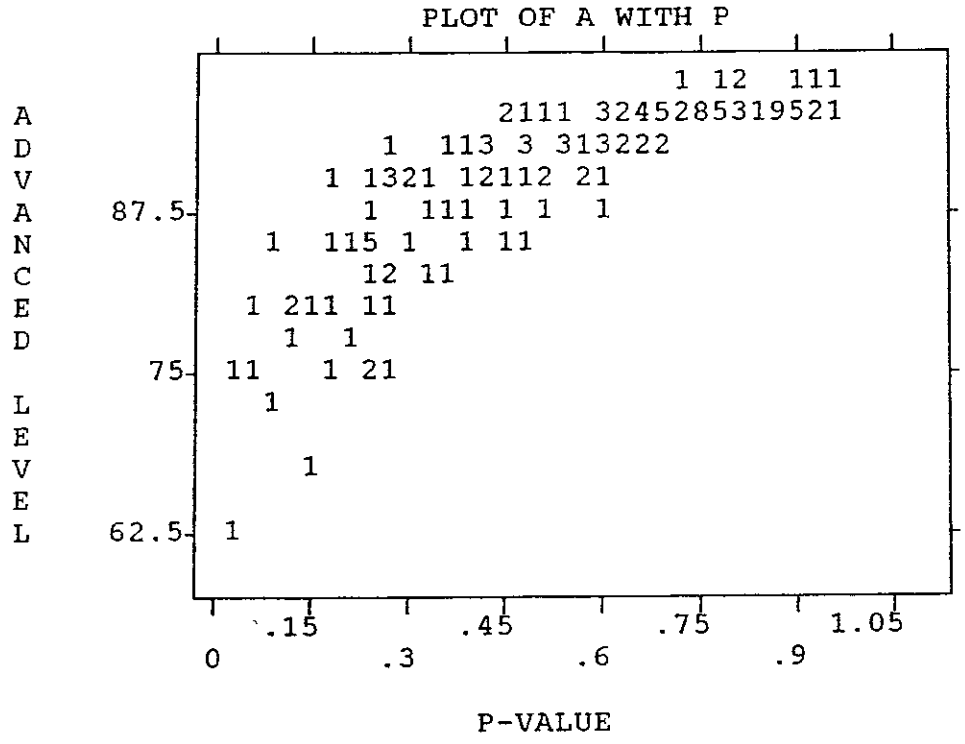
Plot of Adjusted Round 4 Proficient Achievement Levels with Item P-Values for Grade 12



Correlation = .89, Number of items = 144. (correlation of logit transformations of two variables = .91)

Figure 18

Plot of Adjusted Round 4 Advanced Achievement Levels with Item P-Values for Grade 12



Correlation = .84, Number of items = 144. (correlation of logit transformations of two variables = .89)

Figure 19

Regressions of Round 4 Basic Achievement Level on Proportion Correct

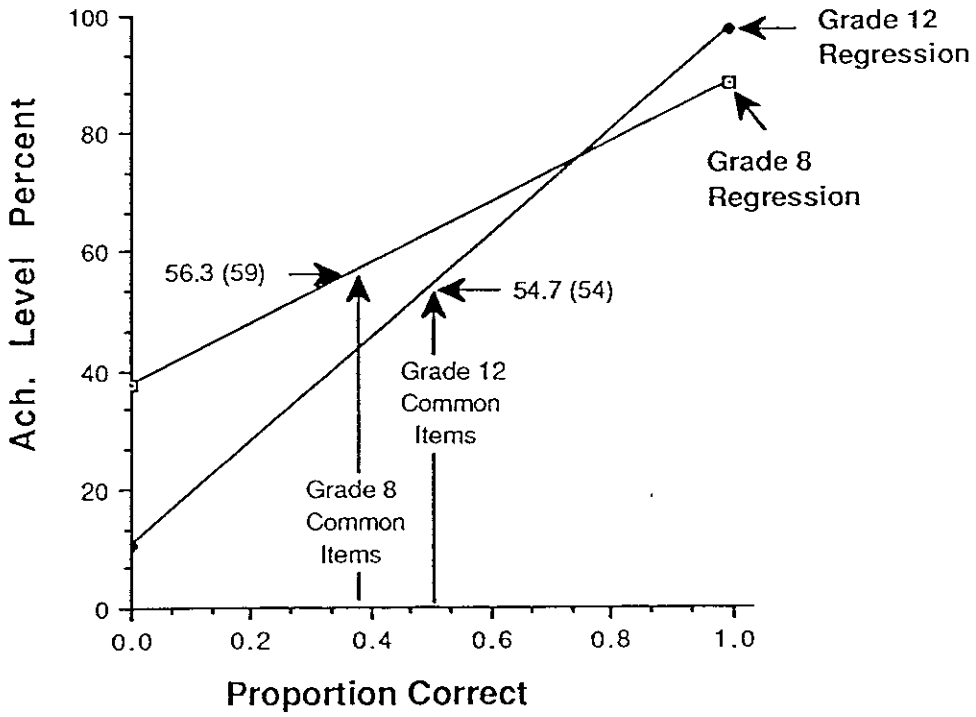


Figure 20

Standard Error Curve for Grade 4 Math Items Used to Set Achievement Levels (Reduced Item Sets)

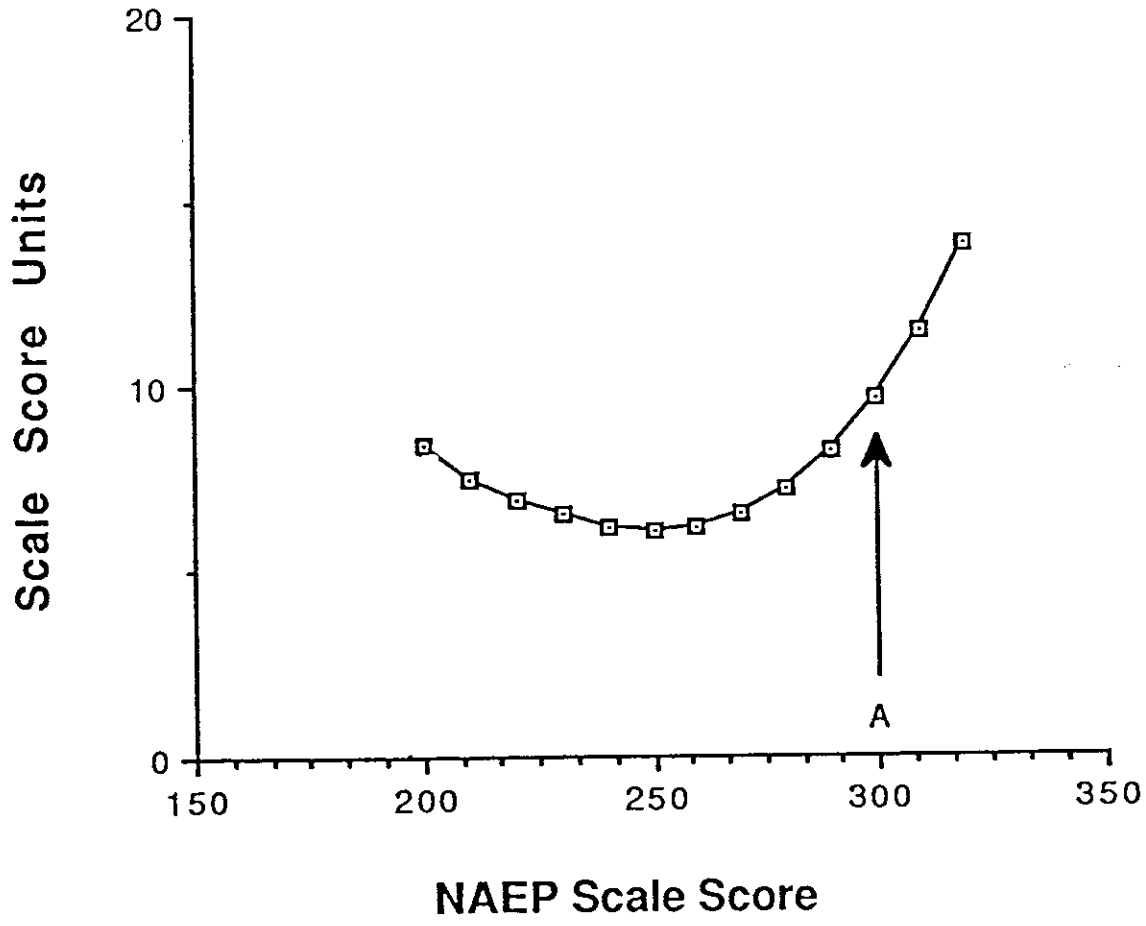


Figure 21

Standard Error Curve for Grade 8 Math Items Used to Set Achievement Levels (Reduced Item Set)

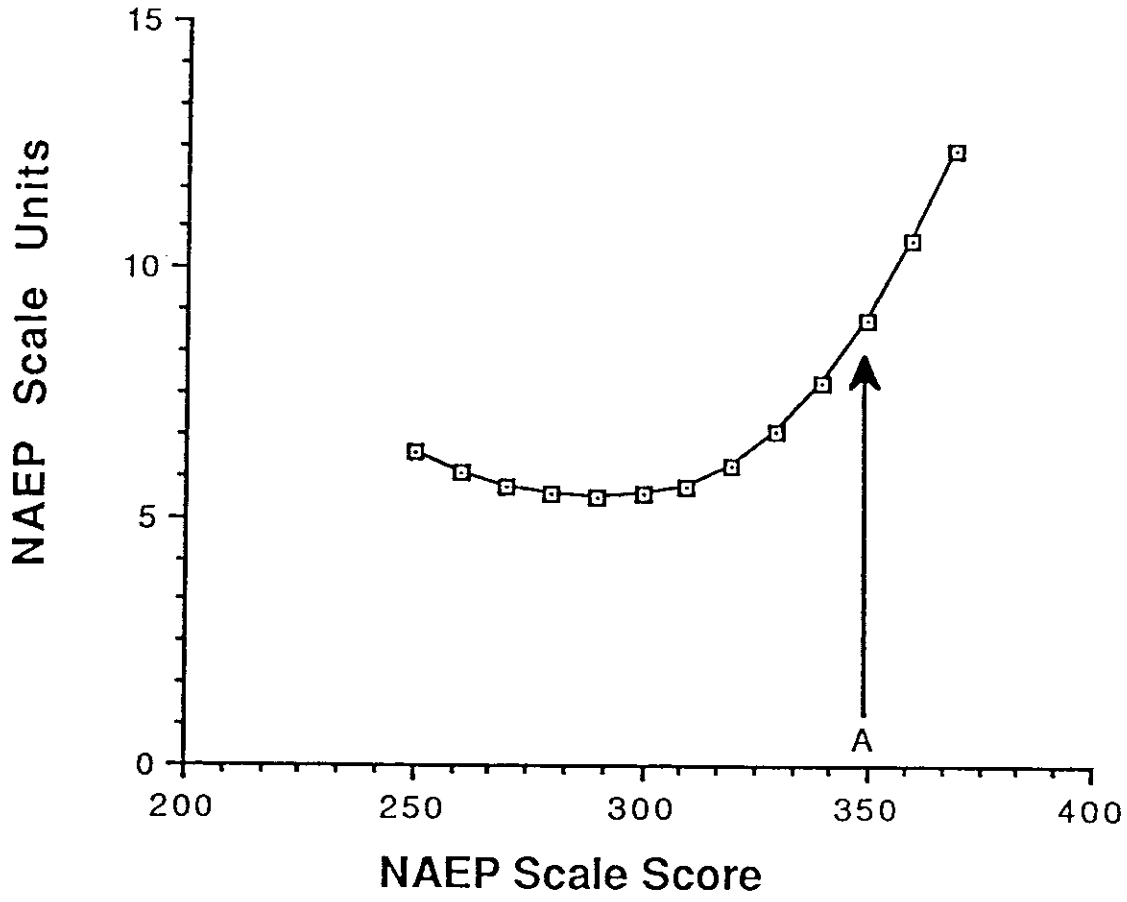


Figure 22

Standard Error Curve for Grade 12 Math Items Used to Set Achievement Levels (Reduced Item Set)