DOCUMENT RESUME

ED 390 910 TM 024 250

AUTHOR Hambleton, Ronald K.; Slater, Sharon
TITLE Using Performance Standards To Report National and State Assessment Data: Are the Reports Understandable and How Can They Be Improved?
PUB DATE Oct 94
NOTE 40p.; Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments (Washington, DC, October 5-7, 1994.) Also published as "Laboratory of Psychometric and Evaluative Research Report No. 271" of the School of Education, University of Massachusetts, Amherst.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment; Elementary Secondary Education; Evaluation Utilization; *Performance; Performance Based Assessment; *Policy Formation; *Research Reports; Scaling; Scores; *Standards; State Programs; Testing Programs; *Test Results
IDENTIFIERS Large Scale Programs; Meaningfulness; National Assessment of Educational Progress; Standard Setting

ABSTRACT

Considerable evidence suggests that policy-makers, educators, the media, and the public do not understand national and state test results. The problems appear to be two-fold: the scales on which scores are reported seem confusing, and the report forms themselves are often too complex for the intended audiences. This paper addresses two topics. The first is to make test-score reporting scales more meaningful for policymakers, educators, and the media. Of particular importance in work on the National Assessment of Educational Progress (NAEP) was the use of performance standards in score reporting. The second topic is the actual report forms that are used to communicate results. Results from a recent interview study with 60 participants using the Executive Summary of the 1992 NAEP Mathematics Assessment were used to highlight problems in score reporting and to suggest guidelines for improvement. The burden is on the reporting agency to ensure that reporting scales are meaningful and that reported scales are valid for the recommended uses. (Contains 3 tables, 4 figures, and 21 references.) (Author/SLD)

Using Performance Standards to Report National and State Assessment Data: Are the Reports Understandable and How Can They Be Improved?

Ronald K. Hambleton and Sharon Slater
University of Massachusetts at Amherst

Abstract

There is considerable evidence available to suggest that policy-makers, educators, the media, and the public do not understand national and state test results. The problems appear to be two-fold: the scales on which scores are reported seem confusing, and the report forms themselves are often too complex for the intended audiences.

This paper was prepared to address two topics. The first is the topic of test score reporting scales and making them more meaningful for policy-makers, educators, and the media. Of special importance in our work was the use of performance standards in score reporting. The second topic is the actual report forms that are used to communicate results to policy-makers, educators, and the public. Some results from a recent interview study with 60 participants using the Executive Summary from the 1992 Mathematics Assessment were used to highlight problems in score reporting and to suggest guidelines for improved score reporting.

Using Performance Standards to Report National and State Assessment Data: Are the Reports Understandable and How Can They Be Improved?[1,2]

Ronald K. Hambleton and Sharon Slater
University of Massachusetts at Amherst

Consider these two quotes from newspaper stories about the 1990 NAEP national and state mathematics results:

> Just one in seven eighth grade students nationwide can exhibit average proficiency in mathematics.

> Standardized tests of student achievement have shown a peculiar quirk for some time now: Every state's kids somehow manage to score above average. Now at least we've got something different - a national math test in which every state's kids scored below average.

The writer of the first story needs a lesson in basic statistics. If some students score below the average then other students must score above it. This same writer also confuses the NAEP scale for reporting proficiency scores with the category or interval on the NAEP scale associated with being "proficient" (other intervals exist too for "below basic", "basic", and "advanced"). These categories or intervals are defined by the performance standards on the NAEP scale.

As for the second quote, again, the writer would benefit from a lesson in statistics. Every student on a NAEP Assessment, or any other test, cannot be below average no matter how poor the overall group performance.

Clearly, both quotes are misstatements of the actual NAEP results. Perhaps they were made to help sell newspapers. A more likely explanation in these two instances is that the writers did

---

[1]Presentation at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, October 5-7, 1994.

[2]Laboratory of Psychometric and Evaluative Research Report No. 271. Amherst, MA: University of Massachusetts, School of Education.

not understand the NAEP reporting scale or the performance standards that were set on the NAEP scale to aid in reporting results. This latter explanation seems plausible as there are still many persons who are unable to distinguish percentages from percentiles, and who think a score of 70 on an IQ test is passing and 100 is perfect. Examples of misinterpretations of educational and psychological test scores abound.

Beyond whatever basic quantitative literacy may be lacking on the part of educators, policy-makers, and the press with respect to understanding assessment results, the fact is that interpreting test scores is always going to be a cognitively challenging activity. Not only do the reporting scales themselves vary from one test to the next, but both measurement and sampling errors must be considered in interpreting scores. Add performance standards to the scale and the task of interpreting scores becomes even more complex.

This paper was prepared to address two topics related to the use of performance standards in score reporting. The first is the topic of test score reporting scales and making them more meaningful for policy-makers, educators, and the media. Of special importance in this paper is the use of performance standards or standards (or achievement levels, as they are called by the National Assessment Governing Board). The second topic is the actual report forms that are used to communicate results to policy-makers, educators, and the public. A few results from a study in which we investigated the understandability of a NAEP

report will be used to illustrate several problems encountered by policy-makers, educators, and the media.

Reporting Scales

What in the world does a NAEP score of 220 mean? This was a common question from policy-makers, educators, and the media in a study we conducted recently using the executive summary of the 1992 NAEP national and state mathematics results. This is a common question from persons attempting to make sense of IQ, SAT, ACT, and NAEP scores, for example. The fact is that people are more familiar with popular ratio scales, such as those used in measuring distance, time, and weight, than they are with educational and psychological test score scales. Even the thermometer scale, which is an equal-interval scale, has meaningful numbers on it to help users understand and interpret temperature scores when they need to. These include 32, 68, as well as daily experiences (such as yesterday's temperature).

In contrast, test scores are much more elusive. Even the popular percent score scale which many persons think they understand is not useful unless (1) the domain of content to which percent scores are referenced is clear, and (2) the method used for selecting assessment items is known.

One solution to the score interpretation problem is simply to interpret the scores in a normative way, i.e., scores obtain meaning or interpretability by being referenced to a well-defined norm group. All of our popular norm-referenced tests use norms to assist in test score interpretations. On the other hand, many state and national assessments are examples of criterion-

referenced assessments, and with these assessments, scores need to be interpreted in relation to content domains, anchor points, and performance standards (Hambleton, 1994).

With NAEP, an arbitrary scale was constructed with scores in theory ranging from 0 to 500 for each subject area. The scale was obtained in, basically, the following way: the distributions of scores from nationally representative samples of grade 4, 8, and 12 students were combined and scaled to a mean of 250 and a standard deviation of about 50 (Beaton & Johnson, 1992). The task then was to facilitate criterion-referenced score interpretations on this scale (see, for example, Phillips, et al., 1993). Placing benchmarks such as grade level means, state means, and performance of various sub-groups of students (such as males, females, Blacks, Hispanics, etc.) is helpful in bringing meaning to the scale, but these benchmarks provide only a norm-referenced basis for score interpretations.

One of the ways of making statistical results more meaningful to intended audiences is to report the results by connecting them to numbers that may be better understood than test scores and test score scales. For example, when the FAA wanted to calm the public's fears recently about flight safety, they reported that, with current safety records, a person could fly everyday for the next 28,000 years without being involved in a serious flight mishap. The connection between accident rates and the number of years of traveling without an accident probably was a meaningful connection for many persons and helped them better understand the current record of flight safety.

Insert Table 1 about here.

Table 1 displays some NAEP scores for students at the 10th, 50th, and 90th percentiles on the 1992 mathematics assessment in grades 4, 8, and 12. One of the reported results was that the average grade 8 student in 1992 performed five points higher (i.e. better) than the average grade 8 student in 1990 (Mullis, Dossey, Owen, & Phillips, 1993). It is clear from Table 1 that the typical student (i.e. the student at the 50th percentile) between fourth and eighth grade gained about 48 points, which converts to about 1.2 points per month of instruction (a gain of 48 points over 40 months of instruction). Recognizing that the growth over the four years is not necessarily linear (see, for example, grade-equivalent scores on standardized achievement tests), it might be said that a gain of five points, is very roughly equivalent to about six months of regular classroom instruction (5 points x 1.2 points gain per month) between grade 4 and grade 8. A five point gain in mathematics achievement for the average student moving between the 4th and 8th grade is very sizable and practically significant and this point would be clear to most persons if the gains were reported in terms of months of instruction required to achieve the gain. Using Table 1, similar interpretations could be set up for low and high achieving students (i.e. students at the 10th and 90th percentiles of the score distribution) between grades 4 and 8, as well as grades 8 and 12.

Reporting score gains in terms of equivalent months of instruction is one convenient way for audiences to have understanding of the meaning of NAEP scores and gains in achievement. Skip Livingston noted in a question to us during the NCES-NAGB conference that it appeared we had simply reinvented the unpopular and commonly misinterpreted grade-equivalent scores. Certainly we are using the grade-equivalent score concept. But, since scores are not being reported for individual students, and given the way in which grade-equivalent scores are used in our approach, most of the well-known shortcomings of grade-equivalent scores do not arise. The main advantage of our approach (i.e. improved communication of the meaning of NAEP scores and gains) seems to far out-weigh any disadvantages of this approach to interpreting NAEP scores.

Other possibilities of considerable promise for criterion-referenced interpretations of scores include anchor points and performance standards (see, Phillips, et al., 1993). Both possibilities, however, have caused controversy and debate in the measurement community (see, Forsyth, 1991; National Academy of Education, 1993; Stufflebeam, Jaeger, & Scriven, 1991).

Both anchor points and performance standards capitalize on the fact that IRT-based scales locate both the assessment material and the examinees on the same reporting scale. Thus, at any particular point (i.e., ability level) of interest the sorts of tasks that examinees at that ability level can handle can be described. And, if of interest, tasks that these examinees cannot handle with some stated degree of accuracy (e.g., 50%

probability of successful completion) can be identified. Descriptions at these points of interest can be developed and exemplary items could be selected, too, that is items could be selected to highlight what examinees at these points of interest might be expected to be able to do (see, Mullis, 1991).

Insert Figure 1 about here.

Figure 1 shows the "item characteristic curves" for two NAEP items (see, Hambleton, Swaminathan, & Rogers, 1991). At any point on the NAEP achievement (i.e. proficiency) scale, the probability of a correct response (i.e. answer) can be determined. Item 2 is the more difficult item since regardless of ability the probability of a correct response to item 2 is lower than item 1. The ability at which an examinee has a 80% probability of success on an item is called the "RP80" for the item. In Figure 1, it can be estimated that the RP80 for item 1 is about 210 and the RP80 for item 2 is about 306. This is known as "item mapping" in that each item in NAEP is located on the NAEP achievement scale according to RP80 values. If 80% probability is defined as the probability at which an examinee can reasonably be expected to know something or be able to do something (and other probabilities have often been used, say 65%, with the corresponding RP65 values) then an examinee with an ability score of (say) 210, could be expected to answer items like item 1 and other items with RP80 values around 210 on a fairly consistent basis (i.e. about 80% of the time). In this

way then, a limited type of criterion-referenced interpretation can be made even though examinees with scores around 210 may never have actually been administered item 1 or other items like it as part of their assessment.

The validity of the criterion-referenced interpretations depends on the extent to which a unidimensional reporting scale fits the data to which it is applied. If a group of examinees scores at (say) 270, then a score of 270 can be made meaningful by describing the contents of items like those with RP80 values around 270. The "item mapping method" is one way to facilitate criterion-referenced interpretations of points on the NAEP scale or any other scale to which items have been referenced. Cautions with this approach have been clearly outlined by Forsythe (1991). One of the main concerns has to do with the nature of the inferences which can legitimately be made from predicted performance on a few test items.

Insert Figure 2 about here.

A variation on the item mapping method is to select arbitrary points on a scale and then to thoroughly describe these points via the knowledge and skills measured by items with RP80 values in the neighborhood of the selected points. In the case of NAEP reporting, arbitrarily selected points have been 150, 200, 250, 300, and 350. Then the item mapping method can be used to select items which can be handled by examinees at those points. For example, using the item characteristic curves

reported in Figure 2, at 200, items such as items 1 and 2 might be selected. At 250, item 4 would be selected. At 300, items 4 and 5 would be selected. At 350, item 6 would be selected. Of course, in practice there may be many items available for making selections to describe the knowledge and skills associated with performance at particular points along the ability scale. With NAEP currently, RP65 values rather than RP80 values are used, and in addition, items which clearly distinguish between anchor points are preferred when describing anchor points. For more details on current practices, readers are referred to Mullis (1991), Phillips et al. (1993) and Beaton and Allen (1992).

The National Assessment Governing Board was not completely happy with the use of arbitrary points (i.e. anchor points) for reporting NAEP results. For one, the points, 200, 250, and 300 became incorrectly associated by the media and policy-makers with the standards of performance demanded of grades 4, 8, and 12 students, respectively. To eliminate the confusion, as well as to respond to the demand from some policy makers and educators for real performance standards on the NAEP scale, NAGB initiated a project to establish performance standards on the 1990 NAEP Mathematics Assessment (Bourque & Garrison, 1991; Hambleton & Bourque, 1991) and has conducted similar projects to set performance standards on NAEP Assessments in 1992 and 1994. The standards have been controversial but that topic will not be taken up here (see, for example, American College Testing, 1993; National Academy of Education, 1993; Stufflebeam, Jaeger, & Scriven, 1991). The important point here is that the performance

standards provide, to the extent that validity evidence supports their use, an additional basis for interpreting scores within a criterion-referenced framework.

Insert Figures 3 and 4 here.

Figure 3 depicts, basically, the way in which performance standards (set on the test score metric, a scale which is more familiar to standard-setting panelists than the NAEP achievement scale) are mapped or placed onto the NAEP achievement scale using the "test characteristic curve" (TCC). (The TCC is a weighted average item characteristic curve for the items which make up the assessment.) With the performance standards for a particular grade on the NAEP achievement scale, these standards can be used to report and interpret the actual performance of the national sample or any subgroup of interest. This situation is represented in Figure 4. With the performance standards in place, the percent of students in each of the performance categories in score distributions of interest can be reported, and the changes in these percents can be monitored over time.

Anchor points and performance standards are placed on an achievement scale to enhance the content meaning of scores and to facilitate meaningful criterion-referenced interpretations of the results (e.g. What percent of grade 4 students in 1992 are able to perform at the proficient level or above?). In NAEP reporting, in recent years, both anchor points (e.g., 150, 200, 250, 300, and 350) and performance standards (e.g., borderline

scores for basic, proficient, and advanced students at grades 4, 8, and 12) have been placed on these NAEP scales. Many states have adopted similar techniques for score reporting.

Performance standards are more problematic than anchor points because they require a fairly elaborate process to establish (e.g. 20 panelists working for five days at a grade level) and validate. At the same time, performance standards appear to be greatly valued by many policy-makers and educators. For example, many state departments of education use performance standards in reporting, and many states involved in the NAEP trial state assessment, have indicated a strong preference for standards-based reporting over the use of anchor points.

Standards-Based Reporting

Performance standards can provide a useful frame-of-reference for interpreting test score data such as NAEP. And, with respect to NAEP, policy-makers, educators, media, and the public need a frame of reference to make sense of the plethora of statistical information coming from 25 years of national assessments. Scaled scores without performance standards (or anchors) would convey little meaning to anyone. But it is not enough to have defensible and valid performance standards. They must be reported and used in ways that interested audiences will understand and interpret correctly (see Wainer, 1992, for examples of problems in reporting data).

Our research described in this portion of the paper was funded by the National Center for Education Statistics NCES) as stimulated by several recent studies conducted on NAEP reports

which found that policy-makers and the media were misinterpreting some of the texts, figures, and tables (Jaeger, 1992; Linn & Dunbar, 1992; Koretz & Deibert, 1993). Our purposes in this study were: (1) to investigate the extent to which NAEP executive summaries were understandable to policy-makers, educators and the media, and to the extent that problems were identified, (2) to offer a set of recommendations for improving performance standard-based reporting practices. Such a study seems essential because there is an unevenness in the measurement literature: there are relatively large amounts of literature on a variety of technical topics such as test development, reliability, validity, standard-setting, and proficiency estimation, but relatively little work has been done on the topic of reporting test score information to communicate effectively with a variety of audiences (for an exception, see Aschbacher & Herman, 1991). More research is needed to provide a basis for the development of guidelines. This study was a modest first step toward the goal of improving test score reporting.

Basic Methodology

The interview used in this study was designed around the Executive Summary of the NAEP 1992 Mathematics Report Card for the Nation and the States (Mullis, et al., 1993). This particular report was chosen because it was relatively brief and could stand alone for policy-makers and educators. Also, the NAEP Executive Summary Reports are well-known and widely distributed (over 100,000 copies of each Executive Summary are produced) to many people involved in various areas of education.

Further, we thought that the NAEP Executive Summary results which included both national and state results would be of interest to the interviewees who were from different areas of the country. Like most executive summaries, this report's format contains tables, charts, and text to present only the major findings of the assessment. For a more in-depth analysis of the NAEP 1992 Mathematics results, readers would need to refer to some of the more comprehensive NAEP reports prepared by NCES.

Our goal in the interviews was to determine just how much of the information reported in the Executive Summary was understandable to the intended audience. We wanted to attempt to pinpoint the aspects of reporting which may be confusing to the readers, and to identify changes in the reporting which the interviewees would like to see.

The 1992 NAEP Mathematics Executive Summary Report consists of six sections that highlight the findings from different aspects of the assessment. For each section, interview questions were designed in an attempt to ascertain the kind of information interviewees were obtaining from the report. Interviewees were asked to read a brief section of the report, and then they were questioned on the general meaning of the text or on the specific meaning of certain phrases. Interviewees also examined tables and charts and were asked to interpret some of the numbers and symbols. Throughout the interviews, we encouraged the interviewees to volunteer their opinions or suggestions. This kind of information helped us gain a general sense of what the

interviewees felt was helpful or harmful to them when trying to understand statistical information.

The 60 participants in the interviews represented a broad audience, similar to the intended audience of the NAEP Executive Summary Reports. We interviewed policy-makers, educators, and people in the media in Massachusetts, Connecticut, Washington, D.C., Louisiana, Kentucky, and New York. We spoke with people at state departments of education, attorneys, directors of companies, state politicians and legislative assistants, school superintendents, education reporters, and directors of public relations. Many of the people we interviewed were prominent individuals in their fields, and most held advanced degrees. Despite this, however, many interviewees had problems reading and interpreting the information they were shown.

Major Findings

The interviewees in this study seemed very interested and willing to participate. For most of them, reports like the Executive Summary were regularly received in their offices. They were eager to help us to determine the extent to which these reports were understandable, and to be involved in the improvement of these reports by offering their opinions.

Despite the fact that the interviewees tried to understand the report, we found that many of them made fundamental mistakes. Nearly all were able to generally understand the text in the report, though many would have liked to see more descriptive information (e.g., definitions of measurement and statistical jargon and concrete examples). The problems in understanding the

text involved the use of statistical jargon in the report. This confused and even intimidated many of the interviewees. Some mentioned that, although they realized that certain terms were important to statisticians, those terms were meaningless to them. After years of seeing these terms in reports, they tended to "glaze over" them.

The tables were more problematic than the text for most of the interviewees. Although most were able to get a general feeling of what the data in the tables meant, many mistakes were made when we asked the interviewees specific questions. The symbols in the tables (e.g., to denote statistical significance) confused some, others just chose to disregard them. For example, interviewees often "eyeballed" the numbers to determine if there was improvement, ignoring the symbols next to the numbers denotating statistical significance. Improvement to these interviewees often meant a numerical increase of any magnitude from one year to the next.

Consider Table 1 from the Executive Summary and reproduced as Table 2 in this paper. Policy-makers, educators, and the media alike indicated several sources of confusion:

1. There were baffled by the reporting of average proficiency scores (few understood the 500 point scale). Also, proficiency as measured by NAEP and reported on the NAEP scale was confused with the category of "proficient students".
2. Interviewees were baffled by the standard error beside each percentage. These were confusing because (1) they

got in the way of reading the percentages, and (2) the footnotes did not clearly explain to the interviewees what a standard error is and how it could be used.

3. The "<" and ">" signs were misunderstood or ignored by most interviewees. Even after reading the footnotes, many interviewees indicated that they were still unclear about the meaning.

4. The most confusing point for interviewees was the reporting of students at or above each proficiency category. Interviewees interpreted these cumulative percents as the percent of students in each proficiency category. Then they were surprised and confused when the sum of percents across any row in Table 2 did not equal 100%. Contributing to the confusion in Table 2 was the presentation of the categories in the reverse order to that which was expected (i.e. Below Basic, Basic, Proficient, and Advanced). This information as presented required reading from right to left instead of the more common left to right. Perhaps only about 10% of the interviewees were able to make the correct interpretations of the percents in Table 1.

5. Footnotes were not always read, and were often misunderstood when they were read.

6. Some interviewees expressed confusion due to variations between the NAEP reports and their own state reports.

Table 3 was prepared to respond to many of the criticisms raised about Table 2 by interviewees in the study. Modest field-testing

during the study indicated that Table 3 was considerably less confusing. A simplified Table 3 may be more useful to intended audiences for the report, but Table 3 may be inconsistent with the reporting requirements of a statistical agency such as NCES.

Insert Tables 2 and 3 about here.

Another common problem for the interviewees was reading the charts. In an assessment of national scope, it is often necessary to include quite a bit of information in each chart. This requires the use of some elegant graphical techniques. This also tends to add to the complexity of the charts. Although these charts are impressive in the NAEP report, to those who could not interpret them, they were intimidating. The unfamiliar chart formats were very difficult for many of the interviewees. Once the charts were explained to them, they understood them, but many commented that they either couldn't have figured the charts out on their own; or more commonly, that they simply would not have the time in a typical day to devote to a report requiring so much study.

The footnotes were of little help in explaining the tables and charts. They were often lengthy and contained statistical explanation that the interviewees did not understand. As an example, the following is a footnote that many of the interviewees found particularly confusing:

> The between state comparisons take into account sampling and measurement error and that each state is being compared with every other state. Significance is determined by an application of the Bonferroni procedure based on 946

comparisons by comparing the difference between the two means with four times the square root of the sum of the squared standard errors.

(Taken from Figure 1, pg. 12 of the Executive Summary of the NAEP 1992 Mathematics Report Card for the Nation and the States.)

The first sentence of this footnote would have been sufficient for the audience we interviewed.

Despite the fact that many of the interviewees made mistakes, their overall reactions to the task were positive. Some were surprised to find that when they took the time to look at the report closely, they could understand more than they expected. Again, most noted that they did not have the time needed to scrutinize these reports until they could understand them. When we apologized to one legislator about the shortage of time we may have allowed for the task, he noted that he had already spent more time with us than he would have spent on his own with the report.

Of those interviewees who had problems, once we explained some of the tables and statistical concepts to them, they found the results easier to understand. Of course, there were a few interviewees who became so frustrated with the report or with themselves that they simply gave up trying to understand it.

Everyone offered helpful and insightful opinions about the report. Some common suggestions were made in these comments about how to make the results in reports like the Executive Summary more accessible to those with little statistical background. A comment made by a couple of interviewees was that the report appeared to be 'written by statisticians, for

statisticians." To remedy this, many suggested removing the statistical jargon. It seems that phrases like "statistically significant" do not hold much meaning for the audience we interviewed, and often only intimidated the readers.

Another suggestion was to simplify the tables by placing the standard errors in an appendix. The lengthy footnotes could also be placed in an appendix for those who are interested. These tended to clutter the appearance of tables. Brief footnotes in layman's terms would be preferred by many interviewees in our study. Also, according to many interviewees, presenting some of the information in simple graphs instead of tables would be better. One reason is that a simple graph can be understood relatively quickly.

It can be seen from some of the comments mentioned above, that most interviewees needed to be able to quickly and easily understand reports. They simply did not have much time or were unwilling to spend much time. Some interviewees would even prefer receiving a more lengthy report, if it were just a bit more clear and easy to understand.

Among our tentative conclusions from the study are the following: (1) there was a considerable amount of misunderstanding about standard-based reporting in the NAEP Mathematics Assessment Executive Summary we studied, (2) improvements will need to include the preparation of substantially more user-friendly reports with considerably simplified figures and tables, and (3) regardless of the technical skills of the audiences, reports ought to be kept

straightforward, short and clear because of the shortage of time persons are likely to spend with these Executive Summaries.

On the basis of our limited and preliminary research, several reporting guidelines for NAEP and state assessments can be offered:

1. Charts, figures, and tables should be understandable without reference to the text. (Readers didn't seem willing to search around the text for interpretations.)
2. Always field-test graphs, figures, and tables on focus groups representing the intended audiences; much can be learned from field-testing report forms. (The situation is analogous to field-testing assessment materials prior to their use. No respectable testing agency would ever administer important tests without first field-testing their material. The same guideline should hold for the design of report forms.)
3. Be sure that charts, figures, and tables can be reproduced and reduced without loss of quality. (This is important because interesting results will be copied and distributed and we have all been forced to look at bad copies at one time or another. Correct interpretations let alone interest can hardly be expected.)
4. Graphs, figures, and tables should be kept relatively simple and straightforward to minimize confusion and shorten the time required by readers to identify the main trends in the data.

Currently, we are preparing a final report of our research in which more details on the research study will be provided along with an expanded set of score reporting guidelines (Hambleton & Slater, 1995).

Conclusions

Standards-based reporting, in principle, provides policy-makers, educators, media, and the public with valuable information. But the burden is on the reporting agency to insure that the reporting scales used are meaningful to the intended audiences and that the reported scores are valid for the recommended uses. At the same time, reporting agencies need to focus considerable attention on the way in which scores are reported to minimize confusion as well as misinterpretations, and to maximize the likelihood that the intended interpretations are made. This will require the adoption and implementation of a set of guidelines for standards-based reporting which include the field-testing of all reports to insure that the reports are being interpreted fully and correctly. Special attention will need to be given to the use of figures and tables, which can convey substantial amounts of data clearly if they are properly designed. Properly designed means that they are clear for the audiences for whom they are intended.

The recently published Adult Literacy Study (Kirsch, et al., 1993), conducted by NCES, Westat, and ETS, appears to have benefitted from some of the earlier evaluations of NAEP reporting and provides some excellent examples of data reporting. A broad program of research involving measurement specialists, graphic

design specialists (see, for example, Cleveland, 1985), and focus groups representing intended audiences for reports, is very much in order to build on some of the successes in reporting represented in the Adult Literacy Study and some of the useful findings reported by Jaeger (1992), Koretz and Deibert (1993) and others. Ways need to be found to balance statistical rigor and accuracy in reporting with the informational needs, time constraints, and quantitative literacy of intended audiences.

These are potentially important times for measurement specialists. Policy makers and the public seem genuinely interested in our assessment results. But without improvements to our scales and reporting forms, no matter how well we construct tests and analyze data, we run the serious risk of being ignored, misunderstood, or judged as irrelevant. The challenge to measurement specialists is clear. We now need to get on with the essential research.

References

Aschbacher, P. R., & Herman, J. L. (1991). Guidelines for effective score reporting (CSE Technical Report No. 326). Los Angeles, CA: UCLA Center for Research on Evaluation, Standards, and Student Testing.

American College Testing. (1993). Setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading, and writing: A technical report on reliability and validity. Iowa City, IA: Author.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. Journal of Educational Statistics, 17(2), 191-204.

Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the National Assessment. Journal of Educational Measurement, 29(2), 163-176.

Bourque, M. L., & Garrison, H. H. (1991). The levels of mathematics achievement volume I: National and state summaries. Washington, DC: National Assessment Governing Board.

Cleveland, W. S. (1985). The elements of graphing data. Monterey, CA: Wadsworth.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? Educational Measurement: Issues and Practice, 10(3), 3-9, 16.

Hambleton, R. K. (1994, April). Scales, scores, and reporting forms to enhance the utility of educational testing. Invited presentation at the meeting of NCME, New Orleans.

Hambleton, R. K., & Bourque, M. L. (1991). The levels of mathematics achievement: Initial performance standards for the 1990 NAEP Mathematics Assessment. Washington, DC: National Assessment Governing Board.

Hambleton, R. K., & Slater, S. (1995). Reporting NAEP results to policy makers and educators: Is the information understood? (Laboratory of Psychometric and Evaluative Research Report No. 271). Amherst, MA: University of Massachusetts, School of Education.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Jaeger, R. (1992). General issues in reporting of the NAEP trial state assessment results. In R. Glaser & R. Linn (Eds.), Assessing student achievement in the states (pp. 107-109). Stanford, CA: National Academy of Education.

Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). Adult literacy in America: A first look at the results of the National Adult Literacy Survey. Washington, DC: U.S. Government Printing Office.

Koretz, D., & Deibert, E. (1993). Interpretations of National Assessment of Educational Progress (NAEP) anchor points and achievement levels by the print media in 1991. Santa Monica, CA: Rand.

Linn, R. L., & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. Journal of Educational Measurement, 29(2), 177-194.

Mullis, V. S. (1991, April). The NAEP scale anchoring process for the 1990 Mathematics Assessment. Paper presented at the meeting of AERA, Chicago.

Mullis, V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1993). Executive summary of the NAEP 1992 mathematics report card for the nation and the states. Washington: United States Department of Education.

National Academy of Education. (1993). A report of the National Academy of Educational Panel on the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels. Stanford, CA: National Academy of Education, Stanford University.

Phillips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., & Barton, P. E. (1993). Interpreting NAEP scales. Washington: United States Department of Education.

Stufflebeam, D. L., Jaeger, R. M., & Scriven, M. (1991). Summative evaluation of the National Assessment Governing Board's inaugural effort to set achievement levels on the National Assessment of Educational Progress. Kalamazoo, MI: Western Michigan University.

Wainer, H. (1992). Understanding graphs and tables. Educational Researcher, 21(1), 14-23.

Table 1

1992 NAEP Mathematics Results

Grade	Percentile Points		
	P_{10}	P_{50}	P_{90}
4	175	220	259
8	220	268	315
12	253	300	343

Table 2

National Overall Average Mathematics Proficiency and Achievement Levels, Grades 4, 8, and 12

Grade	Assessment Year	Average Proficiency	Percentage of Students at or Above Advanced	Proficient	Basic	Percentage Below Basic
4	1992	218(0.7)>	2(0.3)	18(1.0)>	61(1.0)>	39(1.0)<
	1990	213(0.9)	1(0.4)	13(1.1)	54(1.4)	46(1.4)
8	1992	268(0.9)>	4(0.4)	25(1.0)>	63(1.1)>	37(1.1)<
	1990	263(1.3)	2(0.4)	20(1.1)	58(1.4)	42(1.4)
12	1992	299(0.9)>	2(0.3)	16(0.9)	64(1.2)>	36(1.2)<
	1990	294(1.1)	2(0.3)	13(1.0)	59(.5)	41(1.5)

>The value for 1992 was significantly higher than the value for 1990 at about the 95 percent confidence level.

<The value for 1992 was significantly lower than the value for 1990 at about the 95 percent confidence level. The standard errors of the estimated percentages and proficiencies appear in parentheses. It can be said with 95 percent confidence that for each population of interest, the value for the whole population is within plus or minus two standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference.

Table 3

National Overall Average Mathematics Proficiency and Achievement Levels, Grades 4, 8, and 12

Grade	Assessment Year	Average Proficiency	Percentage of Students			
			Below Basic	Basic	Proficient	Advanced
4	1992	218>	39%	43%	16%	2%
	1990	213	46	41	12	1
8	1992	268>	37%	38%	21%	4%
	1990	263	42	38	18	2
12	1992	299>	36%	48%	14%	2%
	1990	294	41	46	11	2

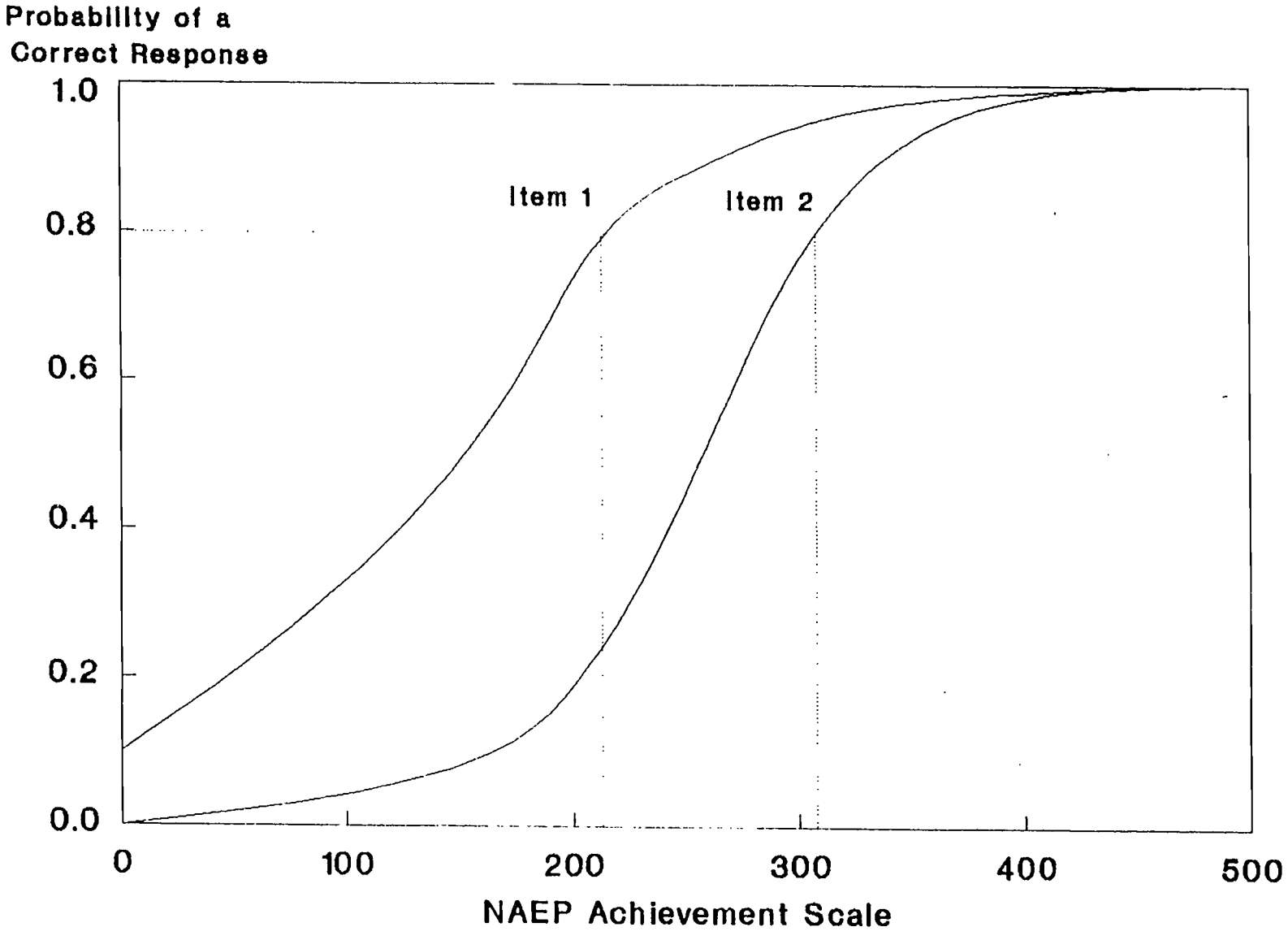
The symbols ">" and "<" are used to highlight differences in the table that are large enough to be real and not due to chance factors such as instability in the information. For example, it can be said that average mathematics performance in Grade 4 in 1992 was higher than in 1990.

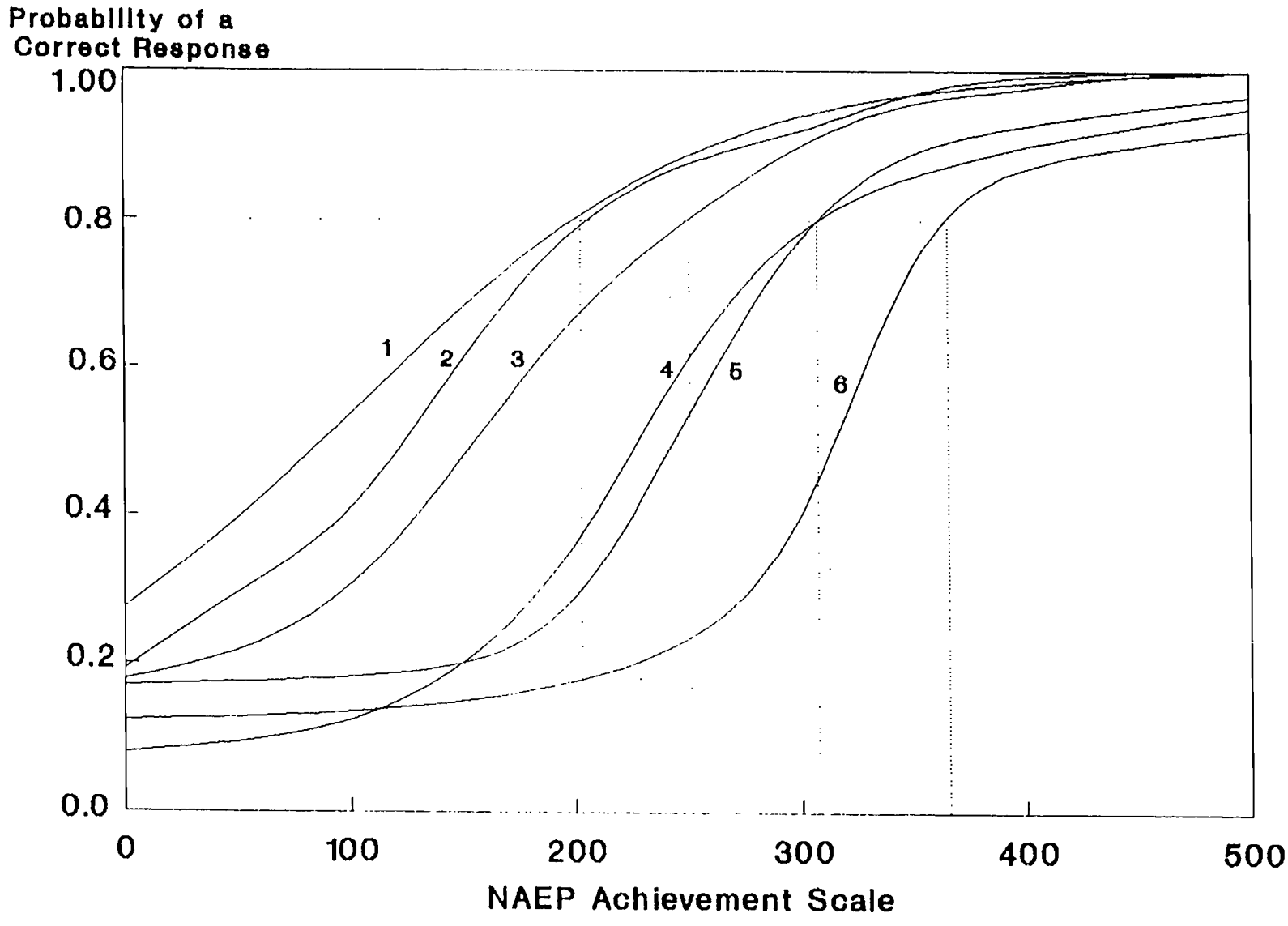
Figure 1. Two test items (or tasks) located on the NAEP achievement scale.

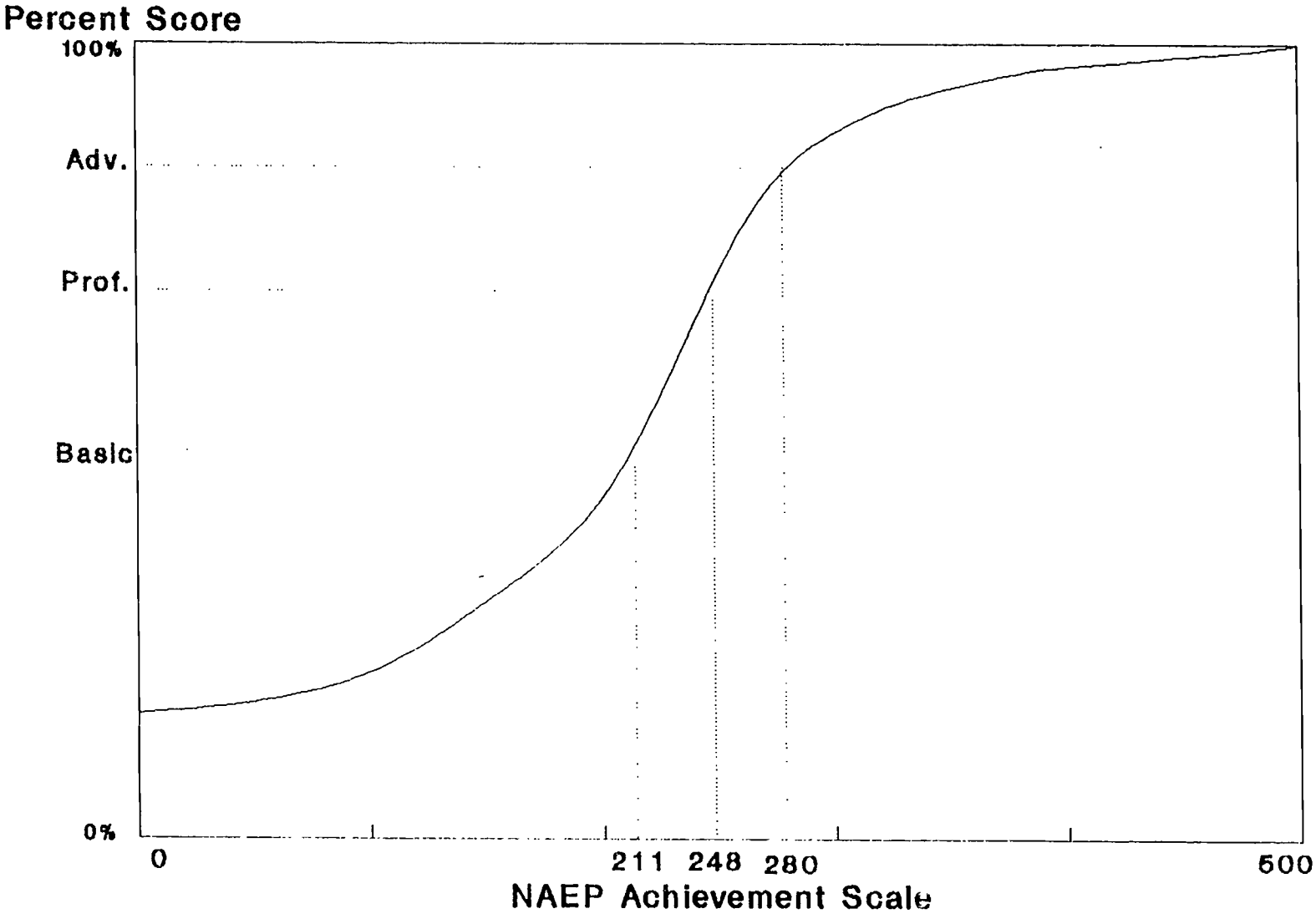
Figure 2. Using anchor points to increase the meaningfulness of NAEP achievement scores.

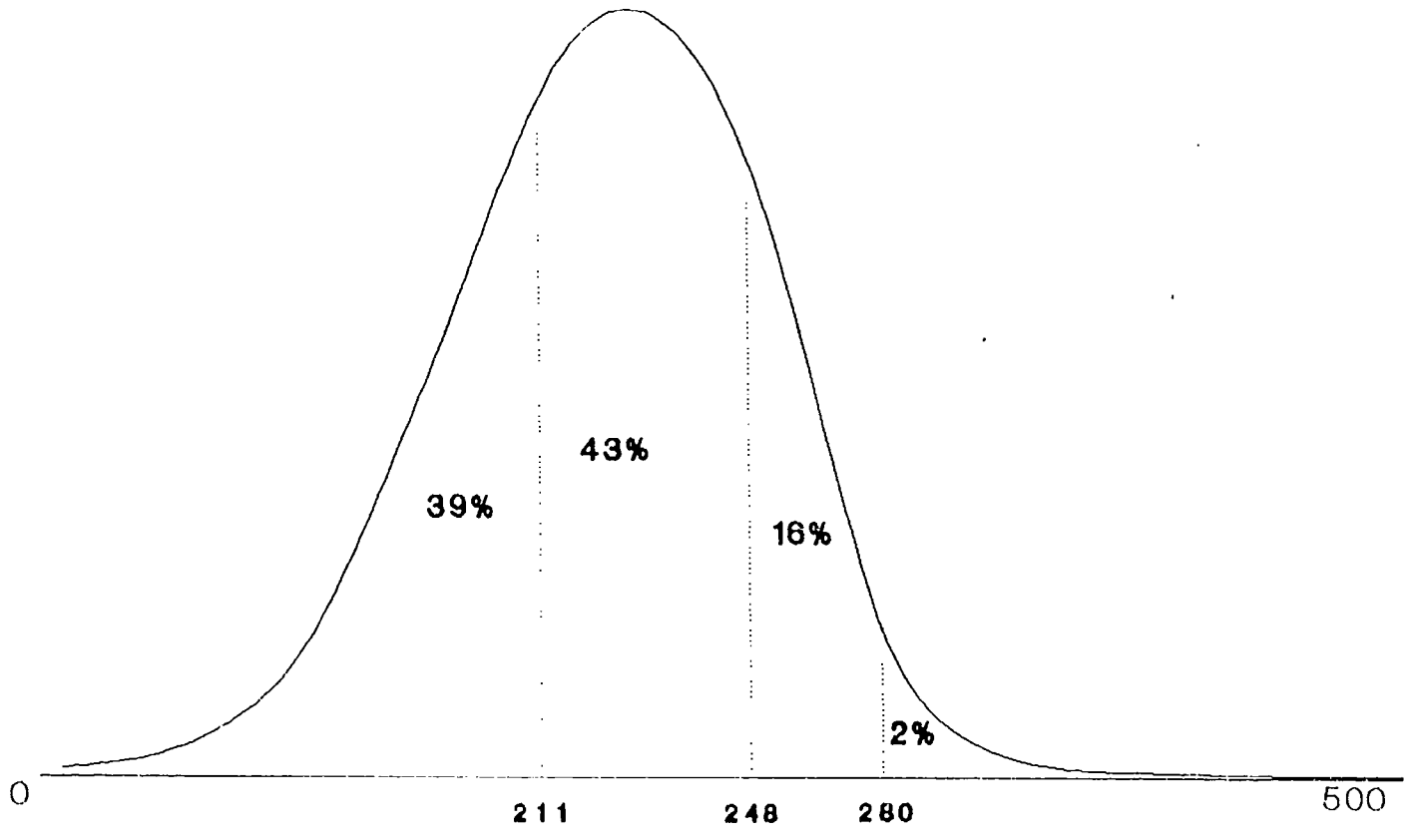
Figure 3. Using performance standards to increase the meaningfulness of NAEP scores.

Figure 4. (Approximate) distribution of 1992 grade 4 NAEP mathematics results.


Probability of a
Correct Response
1.0
0.8
0.6
0.4
0.2
0.0
Item 1
Item 2
0
100
200
300
400
500
NAEP Achievement Scale


Probability of a
Correct Response
1.00
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
NAEP Achievement Scale
1
2
3
4
5
6


Percent Score
100%
Adv.
Prof.
Basic
0%
0
211
248
280
500
NAEP Achievement Scale


39%
43%
16%
2%
0
211
248
280
500
NAEP Achievement Scale