DOCUMENT RESUME

ED 367 683 TM 021 132

AUTHOR Koretz, Daniel; Diebert, Edward
TITLE Interpretations of National Assessment of Educational Progress (NAEP) Anchor Points and Achievement Levels by the Print Media in 1991.
INSTITUTION Rand Corp., Santa Monica, CA. Inst. on Education and Training.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
REPORT NO ISBN-0-8330-1499-4; MR-385-NCES
PUB DATE 93
CONTRACT RS90159001
NOTE 49p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Achievement; Communication (Thought Transfer); Elementary Secondary Education; *Information Dissemination; Journalism; Mathematics Achievement; *Mathematics Tests; National Competency Tests; Newspapers; News Reporting; *News Writing; Periodicals; Scaling; *Test Interpretation; Test Results; Writing for Publication
IDENTIFIERS Accuracy in Media; Anchor Points; *National Assessment of Educational Progress; *Print Media

ABSTRACT

This report reviews the accuracy and reasonableness of statements in the print media about student proficiency on the 1990 National Assessment of Educational Progress (NAEP) in mathematics. It explores the apparent effects of two reporting approaches, the anchor-point method used by the Educational Testing Service and the National Center for Education Statistics since 1984 and the achievement-level (performance standards) method instituted in 1990 by the National Assessment Governing Board. In presenting anchor points and achievement levels the majority of writers incorrectly portrayed performance as discontinuous and ignored the continuum of success and failure. P-values, when provided, were frequently misinterpreted. Only a minority of articles mentioned the judgmental nature of the levels. Use of these metrics appeared to help the press by providing quotable and seemingly clear expressions of test results, but many articles were simplistic or incorrect, and important information often went unreported. Neither method as implemented for the 1990 results was adequate. Some suggestions are made for improved reporting. Two tables present details about the methods. (Contains 40 references.) (SLD)

RAND

Interpretations of National Assessment of Educational Progress (NAEP) Anchor Points and Achievement Levels by the Print Media in 1991

Daniel Koretz, Edward Deibert

Institute on Education and Training

The research described here was sponsored by the National Center for Education Statistics, Contract No. RS90159001, as administered by the U.S. Department of Education. The findings and options expressed in this report do not reflect the position or policies of the National Center for Education Statistics or the U.S. Department of Education.

Library of Congress Cataloging in Publication Data

Koretz, Daniel M.
Interpretations of national assessment of educational progress (NAEP) anchor points and achievement levels by the print media in 1991 / Daniel Koretz, Edward Deibert.
p. cm.
"MR-385-NCES."
"Prepared for the National Center for Education Statistics."
Includes bibliographical references.
ISBN 0-8330-1499-4
1. Mathematics—Study and teaching—United States—Evaluation. 2. Mathematical ability—Testing. 3. National Assessment of Educational Progress (Project). I. Deibert, Edward, 1969– . II. National Center for Education Statistics. III. Title.
QA13.K67 1994
510'.76—dc20 93-44214
CIP

Published 1993 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
To obtain information about RAND studies or to order documents,
call Distribution Services, (310) 451-7002

RAND

Interpretations of National Assessment of Educational Progress (NAEP) Anchor Points and Achievement Levels by the Print Media in 1991

Daniel Koretz, Edward Deibert

Prepared for the
National Center for Educational Statistics
U.S. Department of Education

Institute on Education and Training

The research reported here was conducted for the Technical Review Panel (TRP) of the National Assessment of Educational Progress. The TRP is a joint endeavor of the Center for Research on Evaluation, Standards, and Student Testing (CRESST) at the University of California, Los Angeles; RAND; and the University of Colorado at Boulder.

RAND's Institute on Education and Training conducts policy analysis to help improve education and training for all Americans.

The Institute examines *all* forms of education and training that people may get during their lives. These include formal schooling from preschool through college; employer-provided training (civilian and military); postgraduate education; proprietary trade schools; and the informal learning that occurs in families, in communities, and with exposure to the media. Reexamining the field's most basic premises, the Institute goes beyond the narrow concerns of each component to view the education and training enterprise as a whole. It pays special attention to how the parts of the enterprise affect one another and how they are shaped by the larger environment. The Institute

- Examines the performance of the education and training system
- Analyzes problems and issues raised by economic, demographic, and national security trends
- Evaluates the impact of policies on broad, systemwide concerns
- Helps decisionmakers formulate and implement effective solutions.

To ensure that its research affects policy and practice, the Institute conducts outreach and disseminates findings to policymakers, educators, researchers, and the public. It also trains policy analysts in the field of education.

RAND is a private, nonprofit institution, incorporated in 1948, which engages in nonpartisan research and analysis on problems of national security and the public welfare. The Institute builds on RAND's long tradition—interdisciplinary, empirical research held to the highest standards of quality, objectivity, and independence.

Preface

This report is intended for educators, researchers, and policymakers concerned with the clarity and accuracy with which the National Assessment of Educational Progress (NAEP) communicates information about the proficiency of American students. It describes a review of the accuracy and reasonableness of statements about student proficiency on the 1990 National Assessment in mathematics in reports in the print media. It explores the apparent effects of two reporting approaches—the anchor-point method used by the Educational Testing Service and the National Center for Education Statistics since 1984, and the achievement-level (performance-standards) method instituted in 1990 by the National Assessment Governing Board—on the media reports.

This project was conducted by RAND as part of the work of the NAEP Technical Review Panel (TRP) under contract to the National Center for Education Statistics, U.S. Department of Education. The TRP is a joint endeavor of the University of California, Los Angeles; RAND; and the University of Colorado at Boulder. This study focuses on the uses made of anchor points and achievement levels, the ways in which they were described, and the interpretations of NAEP that were based on those points. The use of simple scaled scores and p-values is also explored—particularly the latter, because of the potential confusion between p-values and the percentages of students reaching anchor points or achievement levels. Another TRP study currently under way is using structured interviews to explore the interpretation of specific NAEP presentations by a sample of policymakers and media writers. A third study explores the validity of the interpretations of the achievement levels presented in National Assessment Governing Board (NAGB) reports of the 1992 mathematics assessment. The opinions presented in this report, however, are solely those of the authors and do not represent the position of RAND or the National Center for Education Statistics.

Contents

Tables

Summary

A primary function of the National Assessment of Educational Progress (NAEP) is to communicate information about the achievement of American students to a wide variety of audiences, including policymakers and the lay public. This difficult goal requires that a large amount of information, some of it arcane, be distilled into simple and clear forms. This study investigates one indication of the effectiveness with which NAEP results have been communicated: the adequacy of reports of the 1990 NAEP mathematics assessment by the print media.

For the past decade, NAEP reporting has centered around scaled scores. One drawback of scaled scores is that lay audiences find them difficult to interpret, because it is not readily apparent how good or bad a score should be considered or how substantial a difference between two groups is. Two methods have been tried to provide NAEP scaled scores with additional meaning. One method uses *anchor points*, which are points arbitrarily chosen from the observed distribution of scores. The percentages of students reaching each of the anchor points are tabulated, and test items (called "anchor items") answered correctly by a large percentage of students at a given anchor point but by a substantially lower percentage at the next-lower anchor point are identified. The characteristics of each set of anchor items are then used to develop verbal descriptions of the performance of students at each level.

The second method, first used in the 1990 mathematics assessment, sets *achievement levels* (labeled Basic, Proficient, and Advanced) according to decisions by a panel of judges about what students in the tested grades *should be able to do*. These levels are then mapped onto the NAEP scale, and the percentages of students who actually did reach those levels are tabulated. As with anchor points, the performance of students reaching the levels is characterized both by verbal descriptions and by displays of illustrative test items.

The anchor-point and achievement-level methods have been the focus of intense debate, but few efforts have been made to explore how they affect the interpretation of NAEP by important audiences, including the press. The 1990 mathematics assessment provided a good opportunity to explore the influence of the anchor-point and achievement-level methods on press accounts because both approaches were used extensively in official reports—the anchor-point method

in reports by the National Center for Education Statistics (NCES; e.g., Mullis, *et al.*, 1991b), and the achievement-level method in reports by the National Assessment Governing Board (NAGB; Bourque and Garrison, 1991) and the National Education Goals Panel (NEGP; 1991). Accordingly, we examined accounts of the 1990 mathematics results in the print media during the seven-month period in 1991 in which the major reports of the assessment were released.

Findings

The anchor-point and achievement-level metrics dominated press coverage of the 1990 mathematics results. During summer 1991, when NCES had released a report using anchor points but no reports had yet been released using achievement levels, virtually all the articles we located used anchor points to report the basic national- and state-level results. Scaled scores were used much less frequently and were rarely used without reference to anchor points. All the articles we located published after the September 1991 release of NAGB and Goals Panel reports (which used achievement levels) relied on achievement levels to report those results. (Both anchor points and achievement levels were used less frequently by the press to report secondary findings, such as sex differences and differences between population groups.)

Descriptions of both anchor points and achievement levels were generally very brief. In describing anchor points, writers often went no further than the capsule description NCES used to label the anchor points in tables. (For example, Level 200, the lowest anchor point, was often described as "simple additive reasoning and problem-solving with whole numbers.") Some writers simplified the descriptions even further, e.g., writing that students at Level 200 "know how to add." Descriptions of the achievement levels were, if anything, simpler yet. For example, many writers used short phrases from the NAGB report, such as "solid academic performance" for the Proficient level. Although both the NCES and NAGB reports provided more substantial descriptions of the knowledge and skills of students at the points or levels, relatively few writers made use of them.

In presenting anchor points and achievement levels, the majority of writers incorrectly portrayed performance as discontinuous. In terms of the skills and knowledge described in NAEP reports, student performance falls on a continuum of success and failure. For example, students just below Level 200 will show considerable success on test items that require "simple additive reasoning"; students just above Level 200 will do only marginally better and will still answer some such items incorrectly. What places a student at an anchor point or achievement level is a particular *rate* of success on items requiring

specific skills or knowledge. The majority of writers made no mention of this continuity of performance. Most used wording that implied that students above an anchor point or achievement level "can do" the things noted in the descriptions and that students below that level cannot do them.

The percentage of students reaching the higher anchor points is lower (often, by a large margin) than the percentage correctly answering the anchor items for those points (called p-values). Although the selection of items to illustrate the achievement levels is less clear-cut, the same disparity often arises with achievement levels. This disparity can lead to a serious underestimate of performance when lay readers incorrectly infer that the percentages of students reaching the levels are the same as the percentage who succeed on the illustrative items. Both the NAGB and NCES reports provided actual p-values (for all students and for students at the anchor points or achievement levels) for illustrative test items. In theory, provision of such information could guard against such a misinterpretation.

The provision of p-values, however, had relatively little effect on the press reports and did not prevent the confusion of p-values with the percentage of students reaching the levels. Relatively few articles presented any illustrative items, and some of the few that did offered no information on the percentage correctly answering them. In the articles following the release of the NAGB and Goals Panel reports, most of those that did present percentages clearly misconstrued the percentage of students reaching the achievement levels as being the p-values for illustrative items.

The achievement levels (unlike the anchor points) reflect judgments about how students should perform, and different panels of judges (or different methods for setting the levels) would likely have produced different standards. Only a small minority of the articles that discussed achievement levels made any mention of the judgmental nature of the levels, and most of those did so only briefly. The implications for the robustness of the levels was not made apparent.

Finally, the use of anchor points and achievement levels resulted in miscellaneous errors of interpretation. Perhaps the most important, in terms of both frequency and significance, was a widespread confusion between anchor points and grade levels of work. The NCES described the anchor points partly in terms of the grades in which the material reflected in the anchor items is commonly introduced or taught. Most writers interpreted that to mean either grade equivalents (e.g., the average performance for each grade) or expected performance for the tested grades. The result was puzzling and even absurd statements—for example, statements that the majority of high school students

perform below the eighth-grade level. Subsequent NAEP reports have not linked the anchor points to grade levels in the same way, however, and more recent press accounts need to be examined to see whether writers have continued to misinterpret the relationships between anchor points or achievement levels and grade levels.

Implications

Anchor points and achievement levels were well received by the press and profoundly influenced press reporting of the 1990 NAEP mathematics assessment. It appears that these metrics help the press by providing quotable, seemingly clear expression of NAEP results. The effects on press coverage, however, are less than satisfactory. Many of the articles reviewed here were simplistic or incorrect, and important information often went unnoted.

No methods of presenting NAEP results can eliminate oversimplification and misinterpretation by the press, but some changes from the strategies used in the 1991 reports might reduce them. The reliance of writers on the anchor points and achievement levels confirms the importance of providing lay audiences with intuitively clear metrics to complement scaled scores. However, the pervasive oversimplification and misinterpretation in the articles reviewed here indicate that neither the anchor-point nor achievement-level method, *as implemented and reported in the 1990 mathematics assessment*, was adequate for this purpose. Clearly, better methods are needed for illustrating the actual profiles of student performance at whatever levels are used in reporting. More effective methods of presentation might entail the following:

- Clear differentiation between actual and expected performance.
- Clearer ways of presenting actual performance on test items used to exemplify the reporting metric. Simply displaying seemingly inconsistent p-values along with the percentages of students reaching various levels on the scale has proven entirely insufficient.
- Explicit and concrete presentation of the continuity of student performance.
- Clear explanation of the role judgment plays in setting standards used for reporting and of the implications of the judgmental nature for proper interpretation of those levels.
- Clear and empirically defensible statements about what students at each of the reporting levels can do on the test.

The effectiveness of these possible changes in reporting must be established empirically. As current reporting methods are refined and new methods are explored, a variety of research methods can—and should—be employed to test their efficacy with NAEP's key audiences.

Acknowledgments

We thank Eric Hamilton for his assistance in coding the articles, and Nancy Rizor for her preparation of the document.

1. Introduction

A primary function of the National Assessment of Educational Progress (NAEP) since its inception has been to communicate information about the achievement of American students—what students know and can do—to the nation at large. This is a difficult goal to meet, because the most essential findings of the assessment must be communicated comprehensibly and accurately to a diverse, nontechnical audience that includes the mass media, policymakers, educators, and interested members of the public. A large and very complex array of information, some technical and arcane, must be greatly simplified and reduced. A central dilemma for the NAEP has therefore been how best to simplify the results so that they are comprehensible but still accurate.

During the more than 20 years that NAEP has been in operation, various methods for communicating its results have been tried. Initially, the reporting procedures focused on the percentage of students answering specific items or types of items correctly; those percentages were accompanied by displays of representative items of each type. Although simple and intuitively clear, this approach poses serious technical problems, particularly for the estimation of trends in achievement. To circumvent those problems, NAEP has relied on scaled scores since the early 1980s. In this method, students' performance is assigned to a numerical scale on the basis of the difficulty as well as the number of items correctly answered. The numerical scale itself is arbitrary; it can be set to have any range the developers of the test choose. For this reason, the technical advantages of scaled scores have a major cost: Points on the scale—and thus scores assigned to students or groups—have no inherent meaning. Their meaning can be discerned only by reference to the distribution of scores along the scale and is therefore not apparent to lay audiences.

Accordingly, two approaches have been used in an attempt to provide intuitive meaning for NAEP scaled scores: the anchor-point method and the achievement-level method. In each method, several points are selected on the NAEP scale (anchor points and achievement levels) and the performance of students reaching those levels is described in nontechnical terms. Although the achievement-level method has other goals as well, both methods can be seen as efforts to provide an intuitively clear simplification of NAEP results while maintaining the advantages of the scaling procedures.

The anchor-point and achievement-level methods have become cornerstones of NAEP reporting,[1] but they have been subjected to substantial criticism. Critics of the anchor-point method have argued that the anchor item descriptions do not consistently match the skills required by the items; that the descriptions have often included unsubstantiated predictive, criterion-referenced interpretations; and that the method leads lay people to underestimate by a large margin the success rates of students on difficult items (e.g., Forsyth, 1991; Koretz, 1989; Linn, 1990). Criticisms of the achievement-level method have been diverse and have focused on the process of setting the levels, the reasonableness of the levels, and the validity of the descriptions and exemplar items used to illustrate them (Burstein, Koretz, Linn, Sugrue, Novak, and Lewis, forthcoming; Linn, Koretz, Baker, and Burstein, 1992; Stufflebeam, Jaeger, and Scriven, 1991; U.S. General Accounting Office, 1992).

Despite such intense debate, few studies have attempted to investigate the final result of these reporting methods: how important audiences actually interpret the findings reported to them.[2] To what degree do readers rely on anchor points and achievement levels in interpreting NAEP results? Are these methods an effective way to concretize and simplify scaled scores and to give them intuitive meaning? Do these methods produce misunderstandings, and, if so, are there patterns in the misunderstandings that offer suggestions for redirection of NAEP reporting?

Accordingly, the Technical Review Panel (TRP) has undertaken a series of studies of the interpretation of NAEP results. This, the first study, investigates the reporting of the results of the 1990 NAEP mathematics assessment by the print media during the latter half of 1991, when four major reports of the 1990 National Assessment in mathematics were released.[3] The articles reviewed were primarily from newspapers, but some newsmagazine articles were located as well.

The following subsection describes the anchor-point and achievement-level methods in more detail. Section 2 describes the methods we used in our literature review. Section 3 describes the results of our analysis of articles that were published in summer 1991, following the release of reports that focused on anchor points. Section 4 discusses articles published in autumn 1991 in response

[1]For more detail about the history of NAEP reporting and the anchor-point and achievement-level methods, see Phillips, Mullis, Bourque, Williams, Hambleton, Owen, and Barton (1993).

[2]One exception is a recent paper by Jaeger (1992).

[3]These reports are described below.

to the release of reports that emphasized achievement levels. The final section, Section 5, discusses conclusions and implications of our findings.

The Anchor-Point and Achievement-Level Methods

Although both the anchor-point and achievement-level methods were designed to give meaning to the NAEP scale by describing performance in specific score ranges, they differed fundamentally in intent and approach.

Setting the Levels

Anchor points are points on the NAEP scale chosen solely on the basis of the observed distribution of scores. They have generally been set 50 points apart, at scores of 200, 250, 300, and 350. (In most subject areas, 50 points is one standard deviation in the total, across-age sample.) Thus, the anchor points are tied neither to *a priori* expectations about student performance nor to the curricula commonly presented in tested grades; nor are they norm-referenced in the conventional sense, because they are not based on information about the *within-grade* distributions of performance.[4]

To characterize performance at each anchor point in the 1990 assessment, students were selected whose scores were within a range of ±12.5 points of each anchor point. Items were then selected based on the performance of students in these groups, using four criteria:

- The p-values (percentage answering each item correctly) within the target group had to be at least .65 in the target group.
- p-values had to be less than .50 in the group below the target group.
- The difference between the p-values in the two groups had to be at least .30.
- These p-values had to be based on samples of at least 100 students (see Mullis, Dossey, Owen, and Phillips, 1991b, Appendix D).

The items, called "anchor items," were then given to expert panels that were asked to characterize the aspects of performance the items illustrated. Their descriptions, in turn, became the basis for the verbal description of the anchor points used in NAEP publications.

[4]The relationship between the across-grade and within-grade variances of scores is not constant across subject areas in NAEP (e.g., Koretz and Lewis, unpublished research); so, anchor points cannot be translated readily into information about the within-grade distributions of performance.

In contrast, *achievement levels* are based on judgments about what student performance *should* be, rather than the current distribution of scores. In 1988, Public Law 100-297 created the National Assessment Governing Board and gave it, among other responsibilities, the mandate to set "appropriate achievement goals" for the subjects and grades tested by the NAEP (see Phillips, *et al.*, 1993, pp. 35 ff.). Those achievement goals have been operationalized by the achievement levels, which were first established for the results of the 1990 mathematics assessment. Levels for the 1992 mathematics and reading assessments were recently established, the former using revised methods.

Three achievement levels—Basic, Proficient, and Advanced—were established for each grade, as follows:

- Panels of judges were given simple definitions of what students should be able to do to be considered as having reached each level.
- Judges were then asked to estimate what proportion of the students who barely qualified for each level could answer specific items correctly.
- These estimated p-values were then mapped onto the NAEP scale.
- On the basis of this process and other information, judges refined and elaborated upon their descriptions of performance at each level.
- Items were then chosen (based on a variety of considerations, including appropriateness of content and actual patterns of performance) to exemplify the levels.

Although this process reflected judges' views about the performance of students just reaching each level, many results were presented in terms of students within the ranges bounded by the levels: below Basic, from Basic to Proficient, from Proficient to Advanced, and exceeding the Advanced level (Bourque and Garrison, 1991; Mullis, Dossey, Owen, and Phillips, 1993).

Descriptions of the 1990 Anchor Points

The anchor points were described in several different ways in the National Center for Education Statistics (NCES) report (see, for example, Mullis, *et al.*, 1991b, Table 1). Uniform brief descriptions of the points were used in tables throughout the report and are summarized in Table 1. The report provided more detailed descriptions of the levels, each two or three paragraphs long, that elaborated on specific skills and knowledge that students at the level showed in various mathematical content areas (Mullis, *et al.*, 1991b, pp. 56–57). Finally, the text of the report offered other characterizations of the anchor points in the course of discussion.

Table 1

Description of Anchor-Point Levels

Level	Description
200	Simple additive reasoning and problem solving with whole numbers
250	Simple multiplicative reasoning and two-step problem solving
300	Reasoning and problem solving involving fractions, decimals, percents, elementary geometry, and simple algebra
350	Reasoning and problem solving involving geometry, algebra, and beginning statistics and probability

Taken together, the brief descriptions, elaborated descriptions, and discussions in the text offered three types of characterizations of student performance at the anchor points: skill-based, grade-based, and predictive. *Skill-based* descriptions are statements about students' knowledge of and skills in mathematics. Both the brief descriptions shown above and the elaborated descriptions were skill-based, and skill-based descriptions dominated the NCES report.

That such skills as the ability to solve two-step problems is a continuum—that is, students just below Level 250 will solve some of such problems, and students just above 250 will solve only marginally more of them—was noted briefly in the NCES report. For the most part, the elaborated descriptions comprise simple statements of what students at a level can do; for example, Level 300 students "can find the perimeters and areas of rectangles, recognize relationships among common units of measure, and use proportional relationships to solve routine problems involving similar triangles and scale drawings" (Mullis, *et al.*, 1991b, p. 57). However, in a number of places, the text of the report addressed the continuity of performance by stating that students at a given level *consistently* succeeded at the tasks noted in the descriptions of the anchor points. For example, in describing the fourth-graders who reached Level 200, the NCES report noted that "approximately 72 percent of the fourth graders demonstrated the ability to *consistently* solve simple addition and subtraction problems with whole numbers." Similarly, eighth-graders who reached Level 250 were described as showing "*consistent* success with multiplication and division of whole numbers" (Mullis, *et al.*, 1991b, p. 7, emphasis added). "Consistent success" could have many meanings, however, and the body of the report offered no explanation of the phrase.[5]

[5]A definition of *consistent* is implied by the criteria for selecting the items used to illustrate the anchor points. Those criteria are contained in an appendix to the report (Mullis, *et al.*, 1991b, Appendix D). However, that appendix does not explicitly clarify what is meant by "consistent success" in the body of the report, and the appendix describes the proportions of students at each level who answered each selected item correctly, not the consistency with which a student at an

The *grade-based* descriptions, which are found in the text of the NCES report, referred to the grade levels at which the material reflected in the anchor items is commonly introduced or taught. (Note that the descriptions did not refer to the grades in which the material is commonly *mastered* by a given proportion of students.) All four of the points were linked to grade levels: "The Panel further characterized Level 200 as material typically covered by the third grade, Level 250 as material generally covered by the fifth grade, Level 300 material as content introduced by the seventh grade, and Level 350 as content generally covered in high-school mathematics courses in preparation for the study of advanced mathematics" (Mullis, *et al.*, 1991b, pp. 6–7). These grade-based descriptions were highly controversial when the NCES report was in review, and NCES removed many references to them in the final materials released in June. A number of references to grade-based descriptions remained in the text of the report, however. (Grade-based descriptions were eliminated entirely in the NCES report of the 1992 assessment; see Mullis, *et al.*, 1993.)

The report sometimes combined the skill-based and grade-based descriptions of the anchor points. For example, in describing the performance of eighth-graders, the report stated that "Virtually all the eighth graders (98 percent) demonstrated a grasp of the third-grade material typified by Level 200—adding and subtracting whole numbers" (Mullis, *et al.*, 1991b, p. 7).

Finally, in text describing the performance of students in grades 8 and 12, the NCES report also described anchor points in *predictive terms*—that is, in terms of students' likelihood of success in later activities. In a description of both eighth- and twelfth-graders, Level 350 was described as "the breadth of understanding necessary to begin the study of relatively advanced mathematics" (Mullis, *et al.*, 1991b, p. 7). For twelfth-graders, the report further elaborated, "these figures show that many students appear to be graduating from high school with little of the mathematics understanding required by the fastest growing occupations or for college work" (Mullis, *et al.*, 1991b, p. 8). NAEP's earlier use of such predictive descriptions of anchor points has been strongly criticized because of the lack of any validating evidence about the actual performance of students in these later activities (Forsyth, 1991; Koretz, 1989).

---

anchor point would be expected to succeed with groups of items categorized in the terms used to describe that anchor point.

Descriptions of the 1990 Achievement Levels

The NAGB report offered a variety of descriptions of the achievement levels. The report provided a table of short definitions of the levels that were used to guide the judges who set the levels, reproduced here in Table 2 (Bourque and Garrison, 1991, p. 5).

In the terminology used above to discuss anchor points, these achievement-level definitions include both grade-based and predictive elements. The *predictive* elements resembled those in the NCES report in referring to readiness for subsequent schooling or employment. Although these definitions also speak in general terms about skills, they are not "skill-based" in the sense that term was used above, because they do not mention any *specific* knowledge or skills that students at each level should (or do) have.

In addition, the NAGB report provided substantial skill-based descriptions of each level at each of the three tested grades (Bourque and Garrison, 1991, Exhibits 1–3, pp. 14–32). Each description began with a short header that turned out to influence press reports substantially: the Basic level was labeled "partial mastery of knowledge and skills"; Proficient was labeled "solid academic performance"; and Advanced was labeled "superior performance." Each level for each grade was further described in terms of "the mathematics knowledge and skills for each level" (Bourque and Garrison, 1991, p. 12). For example, the "understanding of fractions and decimals [of fourth-grade students at the

Table 2

Description of Achievement Levels

Level	Definition
Basic	This level, below proficient, denotes partial mastery of knowledge and skills that are fundamental for proficient work at each grade. For 12th grade, this is higher than minimum competency skills . . . and covers significant elements of standard high-school-level work.
Proficient	This central level represents solid academic performance for each grade tested. . . . It reflects a consensus that students reaching this level have demonstrated competency of challenging subject matter and are well prepared for the next level of schooling. At grade 12, the proficient level encompasses 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.
Advanced	This higher level signifies superior performance beyond proficient. . . . For 12th grade, the advanced level shows readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement.

Advanced level] should extend to a number of representations" (Bourque and Garrison, 1991, p. 14). Each of these one-paragraph descriptions was accompanied by a number of items intended to "illustrate the content of each level" (Bourque and Garrison, 1991, p. 12). Four p-values were presented with each item, for all students and for students at the Basic, Proficient, and Advanced levels.[6]

[6]In this case, students at each level were those whose scaled scores were within a range of ±12.5 points of the level.

2. Methods

Results of the 1990 assessment were released over a period of about half a year. In June 1991, the National Center for Education Statistics released its primary report of the assessment, *The State of Mathematics Achievement* (Mullis, *et al.*, 1991b). This report made extensive use of both scaled scores and anchor points but did not present achievement levels. At the end of September 1991, three additional reports were released, one each by the National Assessment Governing Board, the National Education Goals Panel (NEGP), and NCES. NAGB released its report of the results, *The Levels of Mathematics Achievement* (Bourque and Garrison, 1991), which focused largely on achievement levels and made no use of anchor points. NEGP released its annual report on the status of American students (National Education Goals Panel, 1991). Subtitled *Building a Nation of Learners*, this report included information pertaining to all national education goals, and NAEP results were presented only briefly as one of several indicators addressing Goal 3, student achievement and citizenship. It presented scaled scores in a number of subjects but also gave a prominent place to achievement levels in mathematics, the only subject for which such levels were then available. NEGP used its own terminology for the achievement levels, classifying the Proficient and Advanced levels together as "competent" (National Education Goals Panel, 1991, p. 46). The NEGP report did not mention anchor points. Finally, NCES released an interim summary of a pending report providing a compendium of NAEP trends in a number of subjects over two decades. (The final NCES trends report [Mullis, Dossey, Foertsch, Jones, and Gentile, 1991a] was released in November, also during the period covered by this study.) The trends report used both scaled scores and anchor points extensively but did not use achievement levels.

Our search focused primarily on the June NCES and September NAGB reports because they reported the same data but relied on anchor points and achievement levels, respectively. Four bibliographic sources were searched to locate articles about them: The *National and Regional Newspapers Index* covers 35 daily newspapers from around the country. The *Magazines Index* includes more than 100 diverse periodicals and journals. The *New York Times Index* covers all articles appearing in that newspaper. The *Readers' Guide to Periodical Literature* is a standard, broadly based guide to periodicals and journals. All four were searched for references to the National Assessment of Educational Progress, NAEP, the National Assessment Governing Board, and NAGB.

For the summer articles about anchor points, these four sources were searched for the period June through August 1991. Full text displays were retrieved for all articles that appeared to reference *The State of Mathematics* or the results reported there. This procedure netted 51 pieces: 41 news articles (from 32 different newspapers) and 10 editorials. The number of authors represented by these 41 articles could not be fully ascertained but was sizable. Three authors had more than one article; one had five articles; and two had two each. Eleven articles were given wire-service bylines.[1] All but one of the remaining 21 articles had explicit and unique authorship.

To locate articles about the autumn releases of NAEP data, we searched the same sources using the same criteria but for the period September through December 1991. Because our primary focus was the NAGB report, we retrieved full text listings of all entries that appeared to pertain to that report. Numerous articles and editorials that mentioned NAEP or NAGB but did not include any discussion of student performance were dropped.[2]

A total of 67 articles and editorials located by the search (59 articles and 8 editorials) mentioned some aspect of student performance on the NAEP in mathematics. These 67 pieces appeared in 43 different newspapers and magazines.[3] Of the 59 articles, 26 were attributed to wire services; the remaining 33 were the work of 24 authors. Seven authors had more than one article; with one exception, none had more than two. The one exception, a writer (Mary Ann Roser) who also wrote a number of the anchor-point articles, had credit (alone or with other writers) for five articles.

The great majority of the pieces appeared between September 30—when the NAGB and NEGP reports and the NCES preliminary summary of trends were released—and October 2. A small number appeared over the following week, but only two appeared after October 8 (on October 14 and October 22). Although

---

[1]The authorship of articles given wire-service attribution is not always clear. For example, we located one "wire service" article that contained material identical to some in one of the articles by Mary Ann Roser, the author who was credited with five of the articles we located. We were unable to clarify whether she had written that article as well or provided text through a regional wire-service office.

[2]Two of the pieces dropped from the analysis focused primarily on the issue of achievement levels, even though they did not discuss student performance. Both discussed the criticisms of the achievement levels by Stufflebeam, Jaeger, and Scriven (1991) and the resulting request by the House Education and Labor Committee for an investigation by the General Accounting Office (*Orlando Sentinel*, 1991; *St. Paul Pioneer Press Dispatch*, 1991).

[3]One additional article obtained in the search for autumn articles mentioned the June NCES report in a discussion of a state's decision not to participate in the 1992 NAEP Trial State Assessment but did not make any mention of the autumn NAEP releases. Accordingly, that article was omitted from the set of articles netted by the autumn search.

we extended our search until the end of the year, none of the pieces we located appeared after the release of the final NCES trends report, perhaps because our final filter required reference to the NAGB report.

3. Summer Articles: Anchor Points

The June NCES report presented results in several forms: anchor points, scaled scores, and p-values (the percentages of students correctly answering specific items). Although some results were presented using only one metric, many were presented using more than one, thus presenting writers with a choice about which to use or emphasize.

How Extensively Were Anchor Points Used?

Anchor points dominated the discussion of overall NAEP results in the 51 pieces we located from this period. Anchor points were discussed in 95 percent of the articles and all of the editorials. Scaled scores were used far less. Only about half of the articles and one of the editorials referred to scaled scores, and all of those articles made use of anchor points as well. Half of the 51 pieces used anchor points and made no reference to scaled scores.

These patterns were fairly similar for both national and Trial State Assessment (TSA) results.[1] Thirty-seven of the 41 articles and all of the editorials mentioned the national results. Of those, all but two (96 percent) made use of anchor points, whereas only about one-third (16) made any mention of scaled scores. In only one instance were scaled scores (in this case, the national average) used without reference to anchor points. TSA results were reported nearly as often as national results—in all of the articles and 7 of the editorials. Of 48 articles and editorials presenting TSA results, 71 percent made use of anchor points and 42 percent used scaled scores. Nineteen of these 48 pieces mentioned anchor points but not scaled scores in describing TSA results; only 5 mentioned scaled scores but not anchor points.

In contrast, authors made little use of anchor points in describing population-group (race or ethnicity) or sex differences. Twenty-six of the 51 articles and editorials discussed population-group differences. Of those, only 2 used anchor points, and 10 used scaled scores. The majority of these pieces discussed group differences only in general terms, without using either metric to quantify them. For example: "Asian/Pacific Islanders had the best scores, followed by whites,

[1]Only one of the articles was from a newspaper published in a state that did not participate in the Trial State Assessment.

American Indians, Hispanics and blacks" (Nelis, 1991). Sex differences were described in only 14 cases (about one-fourth of the pieces). Almost all discussed the differences in nonquantified terms, such as "There was virtually no difference between boys and girls in the fourth and eighth grades, but by high school, boys slightly outscored girls" (Roser, 1991c). Two of these pieces reported average scaled scores for males and females, but none made use of anchor points for this purpose.

The reason for the infrequent use of anchor points in descriptions of these group differences is not altogether clear. The relevant tables in the NCES report (e.g., Mullis, *et al.*, 1991b, Table 2.1) present both the mean scaled scores of each group and the percentages of students reaching or exceeding each of the anchor points. One possible explanation is that the descriptions of group differences in the text of the NCES report made little use of anchor points. Indeed, much of the relevant text resembled many of the press accounts in providing general descriptions of group differences, with only occasional references to scaled scores or anchor points to quantify them (e.g., Mullis, *et al.*, 1991b, pp. 12, 82–84). Another possible explanation is that presenting group differences in terms of anchor points requires the use of many numbers.

A four-point scale was constructed to characterize the extensiveness with which anchor points were used.[2] The few articles that made no use of anchor points were assigned a value of 1. A value of 2 denoted articles that used anchor points for only a "minor part" of the discussion. Those articles that made substantial use of anchor points, but not to the extent that they were the primary focus of the discussion, received a 3. Articles in which the discussion was primarily or solely about anchor points were assigned a value of 4. Thirty-eight of the 51 articles and editorials (75 percent) were given ratings of 3 or 4. Eleven pieces received a rating of 2, and a few were rated 1 or did not use anchor points at all. The average for both the 41 articles and the 10 editorials was about 3, with slightly less extensive use of anchor points in the editorials.

How Were Anchor Points Characterized?

As noted above, the NCES report provided skill-based, grade-based, and predictive descriptions of the anchor points. We investigated (1) the frequency with which the writers made use of each of the three types of descriptions; (2) the

[2]These articles often discussed more than NAEP results. For example, some discussed other achievement measures or editorialized about the significance of NAEP results. To score articles on this scale, we considered only the portion of each article devoted to reporting NAEP results.

ways the writers characterized each type when they used it; and (3) the writers' use of test items and p-values.

The Extensiveness of Skill-Based, Grade-Based, and Predictive Descriptions

Even though the NCES report relied more on skill-based descriptions than on either grade-based or predictive definitions, all three types of description were used frequently in press reports of the NAEP results. Skill-based descriptions were used most frequently, but grade-based descriptions were a close second. Of the 39 articles that used anchor points in some way, 92 percent used at least one skill-based definition, and 87 percent used at least one grade-based description. Of the 10 editorials, 3 used skill-based definitions, and 8 used grade-based descriptions. One might speculate that the more norm-referenced character of the grade-based descriptions made them more useful for editorial writers. Finally, despite the relative infrequency with which predictive descriptions were used in the NCES report, they too were used in a sizable proportion of the articles reviewed. About half of the articles (but only 1 of the 10 editorials) used a predictive description as well.[3]

The Use of Skill-Based Descriptions

Skill-based descriptions in the press reports were generally simple. Many stayed close to NCES's wording but used only relatively brief phrases, such as the short descriptions used by NCES to label the levels in tables (e.g., Level 200: "simple additive reasoning and problem solving with whole numbers"; Mullis, *et al.*, 1991b, p. 6). In the majority of the articles, no mention was made of the continuous nature of performance. Instead, performance was treated as if discontinuous, and anchor points were discussed as if they were the points at which students became able to do the various things noted in the descriptions. For example, one writer noted that "the rating system placed students at level 200 if they performed simple addition and problem solving with whole numbers; they reached level 250 if they performed simple multiplication and two-step problem solving" (DeWitt, 1991). In some cases, writers combined this

[3]We located an additional predictive description of the 1990 results in a later article about Massachusetts' decision not to participate in the 1992 NAEP TSA (Ribadeneira, 1991). The reference to the 1990 results was a single sentence saying that the NCES report found "only 5 percent of seniors graduating from high school with enough knowledge to take advanced college courses or handle technological jobs crucial to the nation's economy." Because this article was not primarily about the NAEP results and was outside the time span of the summer search, it was not included in the tabulations reported here.

discontinuousness with further simplification of the wording used in the NCES report, creating striking oversimplifications of the NAEP results. For example, one writer asserted that "72 percent of fourth-graders can add and subtract" (Roser, 1991d). Another stated that "almost all of the 2,500 Colorado students tested—99 percent—knew how to add" (Hernandez, 1991).

The Use of Grade-Based Descriptions

Almost all the writers who used grade-based descriptions misunderstood them to be statements about the *actual* or *expected* performance of typical students in a grade rather than about the typical grade level at which material is introduced or taught. Of the 42 articles and editorials that used a grade-based description, 35 (83 percent) made this error in some form. In some cases, it was not possible to determine whether the writers interpreted the anchor points as actual performance—i.e., as grade equivalents—or as expected levels of performance. For example, one article stated, ". . . most high school seniors perform below the eighth-grade level . . ." (Nelis, 1991).[4] Six articles were more precise, labeling the anchor levels specifically as within-grade average proficiency. For example, one article stated that "students in North Dakota outperformed the 39 other states and territories in the National Assessment of Educational Progress survey, with 24 percent achieving average proficiency [Level 300]" (*Baltimore Evening Sun*, 1991). A few specifically labeled anchor points in terms of expected performance. For example, one writer stated that "even in the best-scoring states only one in four eighth-graders was proficient in skills they are expected to have" (Roser, 1991a); another wrote that "a score of 300 is expected of students at the eighth grade level" (Lue, 1991).

An infrequent but intriguing interpretation was found in three papers that said the NCES report "shows that math scores get worse as students move through school" (Roser, 1991b, 1991c; *St. Paul Pioneer Dispatch*, 1991).[5] This interpretation, as all three articles made clear, stemmed directly from the use of grade-based descriptions of the anchor points in the NCES report. Specifically, 72 percent of fourth-graders reach Level 200, which the NCES report labeled as third-grade material, while only 14 percent of eighth-graders reach Level 300, which the

[4]As this quote indicates, a further source of confusion was that the grade levels noted in the anchor level descriptions were not those in which the NAEP was administered. Thus, anchor Level 200 was described in terms of material presented in the second grade, and anchor Level 300 was described in terms of material introduced in the seventh grade, but some writers presented these levels as performance typical of the third and eighth grades, presumably because NAEP results were available for those grades.

[5]The article in the *St. Paul Pioneer Dispatch*, labeled "from wire services," included verbatim quotes from Roser (1991c) and may have been written by her as well.

NCES report labeled as seventh-grade material (Mullis, *et al.*, 1991b, Table 1, p. 6). This interpretation is entirely inconsistent with the NAGB achievement levels, which were designed to express performance in terms of grade-level expectations. NAGB found roughly equal proportions of students reaching the two lower achievement levels (Basic and Proficient) in each grade, while the proportion reaching the Advanced level is smallest in grade 4 and largest in grade 12.

The Use of Predictive Descriptions

Of those articles that used predictive descriptions, the majority followed the wording of the report reasonably closely. Thus, one article included this about the performance of high school seniors on the NAEP: "... many are poorly equipped to work in high-tech industries or to study math at the college level" (Eskey, 1991). A few of the articles, however, further simplified these statements and equated performance at an anchor level to being prepared for college. For example, one article included the following: "The numbers represent levels of math achievement: 150—basic arithmetic; 250—simple problem solving; 350—college ready" (*USA Today*, 1991).

How Were Items and p-Values Used?

As noted earlier, the anchor-point method has been criticized because it has led some observers to underestimate the success rate of students on anchor items and thus to underestimate the overall level of achievement of American students (e.g., Linn, 1990). The proportion of students reaching the higher anchor levels is necessarily considerably lower than the proportion of students correctly answering the items used to anchor that level.[6] This discrepancy can be large; for example, in the 1986 science assessment, 6 percent of 17-year-olds scored above anchor point 350, but the p-values for the eight items used to anchor that level ranged from .25 to .51, with a mean of .39 (Koretz and Lewis, unpublished research). Many observers, however, mistakenly assumed that the proportion of students reaching an anchor point was the same as the proportion of students correctly answering the anchor items.

To address this problem, the 1991 NCES report included p-values, both overall and for students at each anchor point, for a variety of anchor items (Mullis, *et al.*,

[6]This is because some of the many students below the anchor point also answer each anchor item correctly. For further explanation, see Linn, 1990.

1991b, pp. 60 ff.) The articles surveyed here provide an opportunity to see how writers used those p-values and whether they interpreted them appropriately.

The presentation of items and p-values in the NCES report had only a minor effect on the reporting of NAEP results by the press. To begin with, only a modest number of articles included any sample NAEP items at all, with or without p-values. Only about one-fourth of the articles (11 of 41) and only 1 of the 10 editorials included any sample test items. Other articles relied on other means, such as the verbal descriptions of anchor points, to describe for readers what students can do.

Moreover, even among the articles that presented sample items, p-values were rarely mentioned, and the disparities between p-values and the percentages of students reaching the anchor points were not addressed. Seven of these 11 articles included a set of sample problems exemplifying certain anchor levels but did not include p-values for them. One article, for example, stated: "Examples of 300-level problems that many eighth-graders did not answer correctly include computing the combined weight of 50 tomatoes that averaged 2.36 pounds; identifying a decimal between .07 and .08; and converting 3 and 3/10ths, expressed as a fraction, into a decimal" (Cooper, 1991). (The proportions of students who answered these questions correctly were 48 percent, 54 percent, and 50 percent, respectively.) The four articles that presented actual p-values used them correctly but did not attempt to reconcile the differences between them and the percentages of students reaching the anchor points. The sole editorial that presented supposed p-values appears to have confused them with the proportion of students reaching the anchor points. For example, the editorial states "... 7th Grade Problems: 14 percent of eighth-graders and 46 percent of twelfth-graders answered correctly questions at the seventh-grade level." This is then directly followed by a sample question: "What is the least whole number x for which $2x > 11$?" (*Seattle Times*, 1991). The actual p-values for this item, however, were 45 percent in grade 8 and 65 percent in grade 12.

Anchor Points and the Range of Scaled Scores

In the course of reviewing the summer articles, we noticed an unexpected misinterpretation of the NAEP scales that resulted from the use of anchor points. Four articles mistakenly concluded that the anchor points discussed in the NCES report defined the total range of scaled scores. Although the number of articles in which we located this error was small, it was prominently displayed in a front-page article in *The New York Times* (DeWitt, 1991).

4. Autumn Articles: Achievement Levels

Of the 67 relevant articles and editorials located for this period, 4 discussed student performance but presented so little detail that they were not usable for further analysis. For example, some simply reported that student performance had been increasing and had returned to the levels of 20 years ago (e.g., Cohen, 1991; White, 1991). All the tabulations reported below pertain to the 63 remaining articles and editorials.

The frequency with which the writers made use of the three reports is not entirely clear, but it appeared as though the NAGB report was cited in all of them, and the NCES and National Educational Goals Panel reports in about three-fourths (70 percent and 73 percent, respectively). The reason for this uncertainty is that some of the writers did not include explicit references to the reports. Because of the overlap among the reports and the ambiguity of some of the articles, it was not always clear what sources these authors had relied upon.

How Extensively Were Achievement Levels Used?

Without exception, all of the remaining 63 pieces made at least some use of achievement levels in their reporting of NAEP results, but the writers used achievement levels more frequently to report some results than others. All but one of the pieces (62 of 63) reported national results; of those, all but one used achievement levels. Only about half of the pieces (33 of 63) reported state-level TSA results, but in those cases, as well, the use of achievement levels was nearly universal (32 of the 33 pieces). Achievement levels were used relatively infrequently, however, in reporting other differences. For example, 28 of the 63 pieces reported population-group differences; of those, only 8 (29 percent) used achievement levels in reporting those results. Ten pieces reported sex differences, but only 3 used achievement levels to do so. Fifteen (one-fourth) of the pieces reported other group differences, such as differences among regions or socioeconomic groups; 8 of those 15 pieces used achievement levels to describe those results.

Note that the use of achievement levels to describe population-group and sex differences, while infrequent in comparison with the descriptions of national results, was a bit more frequent than the use of anchor points to describe these group differences in the summer articles. This disparity may have occurred

because achievement levels are more central to the descriptions of group differences in the NAGB report (e.g., Bourque and Garrison, 1991, pp. 35 ff.) than in the NCES report (e.g., Mullis, *et al.*, 1991b, pp. 82 ff.).

How Were Achievement Levels Characterized?

We explored four aspects of the writers' portrayals of achievement levels: (1) whether they presented the achievement levels as descriptions of what students can do; (2) the extent to which they relied on predictive definitions of the levels; (3) the amount of detail in their descriptions of the levels; and (4) whether they presented the levels as judgmental.

What Students "Can Do" Versus "Should Be Able to Do"

In recent months, a controversy has arisen about the use of achievement levels to describe what students at various levels actually can do on the NAEP. In contrast to the achievement levels, the anchor points are unambiguous in this respect. The anchor points themselves are taken from the actual distribution of student performance, and anchor items are selected entirely on the basis of the performance of students whose scores are within a given range of adjacent anchor points. Although there is room to question whether judges do an adequate job of characterizing the anchor items verbally, the method is clearly grounded in actual student performance and is designed to characterize it. In contrast, the achievement levels are intended to reflect judgments about what students *should be able to do*—i.e., about the level of performance that is considered acceptable for students in the tested grades.

Even though the achievement levels are intended to reflect judgments about what students should be able to do, in one sense they also are statements about what students can do. The complex process of setting the achievement levels has two basic stages. In the first stage, judges decide what students should be able to do; for example, they decide what levels of performance represent "solid academic performance for each grade" (Bourque and Garrison, 1991, p. 5). In the second stage, these levels are mapped to the NAEP scale, and the NAEP is then tabulated to show how many students *do* reach those levels. For example, the NAGB report stated that "NAEP achievement levels are standards of performance that prescribe what students at each grade *should know* and be able to do based on the NAEP assessment—and such standards allow the estimation of how many American students *have reached those levels*" (Bourque and Garrison, 1991, pp. 11–12, emphasis added). Although the word *should* appears in some of

the descriptions of achievement levels, the NAGB report focuses primarily on characterizations of student performance, not expectations. Thus it seems very likely that readers will interpret the results primarily as statements about actual student performance.

In response to this debate, the NAEP Technical Review Panel explored the extent to which the published descriptions of the 1992 mathematics achievement levels accurately describe the performance of students in the score ranges defined by the levels. This study showed that the achievement-level descriptions are not in fact accurate descriptions of what students can do. Some of the attributes in the descriptions of the achievement levels are not assessed by the NAEP, and thus NAEP provides no evidence about whether students actually can do what is described. The items used to exemplify the levels do not consistently characterize the groups or differentiate among them. Items selected from the NAEP to match the descriptions do not consistently show reasonably high rates of success in the relevant groups of students and do not consistently differentiate among groups. Finally, the items that do differentiate among the groups defined by the levels suggest attributes not mentioned in the descriptions (Burstein, Koretz, Linn, Sugrue, Novak, Lewis, and Baker, 1993).

Although the most recent controversy about the proper interpretation of the achievement levels has focused on the 1992 levels (which were created and described somewhat differently from the 1990 levels), the articles about the 1990 levels reviewed here provide the only opportunity to date to ascertain how the media interpret the achievement-level descriptions. Accordingly, we attempted to ascertain whether the reviewed pieces presented the achievement levels as descriptions of what students can do.

The writers in our sample clearly interpreted the achievement levels as statements about what students actually can do. We concluded that 45 pieces (71 percent) included detailed enough discussion of the levels that the "can versus should" question could be addressed sensibly. In all but one of those 45 pieces, writers presented the achievement levels as statements about what students can do. Four of these articles also made reference to what students should be able to do, but only one article made reference to what students should be able to do without also referring to what students can do.

The Use of Predictive Descriptions

Surprisingly, the autumn articles we reviewed made very little use of predictive descriptions—far less than the summer articles about anchor points. In only 4 of the 63 pieces analyzed were one or more predictive definitions clearly used. One

might speculate that the basic theme of the achievement levels—the proportion of students meeting expectations for their grades—is so clear to lay audiences that writers perceived less need to use the predictive descriptions.

What Detail Did the Writers Provide?

Most often, the writers of the autumn articles and editorials provided only very brief descriptions of the achievement levels. Fifty-one (81 percent) of the 63 pieces used one or more of the labels (Basic, Proficient, and Advanced). Thirteen (21 percent) followed the lead of the Goals Panel and labeled the Proficient "competent." Four used one or more of the brief headers that NAGB used in its detailed descriptions of the levels ("superior performance" for Advanced, "solid academic performance" for Proficient, and "partial mastery of knowledge and skills" for Basic).

Only 12 (19 percent) of the 63 articles and editorials provided additional detail. Typically, these articles drew on the more detailed, content- and skill-based descriptions provided in the NAGB report (Bourque and Garrison, 1991, pp. 14 ff.). For example, one writer (Farmer, 1991) stated that

> Students at [the proficient level] should be able to solve problems requiring decimals and proportions with and without a calculator. Students at the advanced level can solve complex problems involving elementary concepts of probability and can apply basic geometric properties.[1]

Relatively few articles and editorials (16 out of 63) made explicit reference to grade-level work. One particular phrase—"can tackle solid grade level work"—was used by the majority of writers who referred to grade levels. It is striking that grade levels were mentioned less frequently in describing the achievement levels than in articles describing anchor points, given that NCES did not mention grade-level performance in its descriptions of anchor points and placed relatively little emphasis on the grade levels at which curricular material is introduced. However, this omssion may be because grade levels are so central to the definitions of achievement levels that they were made apparent by context rather than explicit reference in the articles that described them.

A very few writers deviated from the NAGB and NEGP descriptions. One writer relabeled the levels, and another referred to the Proficient level as "the national standard" (Asayesh, 1991).

---

[1]Note that this article is one of the few that described the levels in terms of what students *should be able* to do as well as what they *can* do.

Was Performance Presented as Discontinuous?

As in the case of the summer articles, the majority of the autumn articles used wording that implied incorrectly that student performance is discontinuous. Of the 63 pieces, 41 provided enough detail to permit us to address this question. Of those 41 pieces, 40 used terminology suggesting discontinuous performance.

Were the Achievement Levels Described as Judgmental?

The achievement levels represent *judgments* about the levels of performance that students should reach. As such, they are not absolute; different panels of judges, or similar panels given different instructions, would likely have pegged the levels at different points on the NAEP scale. Indeed, variation among panels of judges was one basis for criticism of the 1990 achievement levels (Linn, Koretz, Baker, and Burstein, 1992; Stufflebeam, Jaeger, and Scriven, 1991), and the 1992 judges set levels at somewhat different points than did the 1990 judges. A critical question, then, is whether the key audiences understand that the levels are judgmental and that reasonable people could establish a variety of substantially different levels.

Only a minority of the autumn articles and editorials mentioned the judgmental character of the achievement levels. We excluded 6 pieces in which the discussion of achievement levels was not clear enough to allow us to categorize the presentation as absolute or judgmental. Of the remaining 57 pieces, only 16 (28 percent) indicated that the achievement levels represented judgments, and some of those did so only with a phrase or two. Because of the typically cursory discussion of this point, however, the categorization of articles was not always clear-cut, and these findings should be taken as rough estimates.

How Were Items and p-Values Used?

Like the summer articles about anchor points, the autumn articles about achievement levels made relatively scant use of items and p-values, even though the NAGB report displayed a substantial number of illustrative items with p-values for all students and for students at each level. Moreover, most of the writers who did use specific items explicitly or implicitly confused p-values with the percentages of students reaching the achievement levels.

Only 14 articles or editorials (22 percent) referred to specific items presented in the NAGB report. (Eight additional articles referred only to specific items or tasks from other assessments, such as a previous Educational Testing Service

[ETS] assessment of literacy.) Only a single article correctly reported p-values (Hildebrand, 1991). One presented sample items along with the percentages of students reaching each achievement level but did not indicate which levels the items exemplified (Adams, 1991). Another article provided a purported p-value that appeared to be neither a p-value nor the percentage of students reaching the relevant level. The remaining 11 articles all confused p-values and the percentage of students reaching the achievement levels. In some instances, the error was explicit: Readers were given the percentages reaching achievement levels but were explicitly told that they were the percentages of students correctly answering specific items (e.g., *San Jose Mercury News*, 1991; *Richmond News Leader*, 1991). In other cases, the error was less explicit: Readers were told the proportion of students reaching (or failing to reach) a given level and were then told that students in that category could or should be able to answer one or more specific items correctly. For example, one writer stated: "'What is the value of $n + 5$ when $n = 3$?' was considered a 'basic' math question on the 1990 test given by [NAEP]. For most, it was basic. Substitute 3 for n and the answer becomes 8. Yet nearly 57% of N.C. 8th graders who took this exam couldn't correctly answer questions of this difficulty" (O'Brien, 1991).

5. Conclusions

Anchor points and achievement levels clearly struck a responsive chord among the press writers whose work we reviewed. Almost all the writers relied heavily on these metrics in reporting the basic national- and state-level results of the NAEP. These metrics appear to have met a need for simple expressions of NAEP results that appear intuitively clear and are eminently quotable.

At the same time, neither metric had entirely satisfactory effects on reporting by the press. Press accounts were often inadequate or even were simply wrong. In this section, we review the patterns that appeared in the press reports and, in particular, the weaknesses in some of them. We then suggest possible ways to strengthen the reporting of NAEP results.

Patterns in the Press Reports

One of the most consistent but least surprising of the patterns in the articles reviewed here is the simplicity of many of the presentations of NAEP results. Many articles were short, and some longer articles only briefly mentioned actual NAEP results. Moreover, some of the longer articles provided many specifics—for example, lists of mean scores from all 37 states that participated in the Trial State Assessment—rather than a rich discussion of any particular results.

A particularly important aspect of this simplicity was the typically sparse descriptions of both anchor points and achievement levels. In the autumn articles, for example, few writers provided description of the achievement levels beyond the labels used by NAGB and NEGP (Basic, Proficient, Competent, and Advanced) or the two- and three-word descriptions, such as "solid academic performance," provided by NAGB. In some cases, writers even simplified the NAGB and NCES descriptions further, making statements such as "X percent of students can add."

Moreover, relatively few of the articles showed any understanding of the continuity of performance or a clear notion of how to use anchor points or achievement levels to place students on that continuum. Indeed, one could conclude that an unintended effect of the anchor-point and achievement-level methods was to encourage the misconception of performance as discontinuous. (It would be difficult to interpret simple scaled scores, presented alone, as

discontinuous.) Many writers interpreted anchor points as the points at which students can do the things noted in the anchor-point descriptions, rather than the points on a continuum indicating that the rate of success on those tasks reached a certain level. Similarly, relatively few writers made any use of the p-values provided in the NAGB and NCES reports, and virtually none contrasted those p-values with the percentages of students reaching anchor points or achievement levels.

Consistent with the typically simple presentation of NAEP results, substantively important details that were not made prominent in the reports often received no discussion. For example, in keeping with the discontinuous portrayal of student performance in many articles, the frequent but passing references in the NCES report to "consistent" success on given types of items had little or no impact on press reports. Similarly, references to what students "should be able to do" in the longer NAGB descriptions of achievement levels were not reflected in most of the autumn articles, and none drew a clear contrast between that and what students can do.

The articles reviewed here showed a tendency to rely on familiar metrics or even to translate novel metrics into conventional, familiar ones. One example of the former is the use of (unvalidated) predictive descriptions of the anchor points in some of the summer articles. Although such descriptions were used in only a minority of the articles, even this infrequent use seems disproportionate to the relatively minor emphasis predictive descriptions were given in the NCES report. A striking instance of translation of metrics was the frequent conversion of anchor points into grade equivalents or expectations for specific grades (between which many authors drew no distinction). Although the NCES report provided ample encouragement to make this translation by making frequent references to the grades in which material is introduced or taught, many of the press reports went well beyond the NCES report in this respect.

Finally, most of the autumn articles made no reference to the judgmental basis of the achievement levels. This omission is not surprising, given the presentation of the levels in the NAGB report, which provided a brief summary of the judgment process but did not clarify the implications of this process for the robustness and appropriate interpretation of the achievement levels. For example, the NAGB report concluded a summary of the judgment process with the simple statement that "NAEP achievement levels are standards of performance that prescribe what students at each grade should know and be able to do based on the NAEP assessment," with no caveat indicating that these standards are to some extent a function of the judges selected and the methods used (Bourque and Garrison,

1991, pp. 11–12). Nonetheless, it is an important shortcoming of the media portrayal of the levels.

Recommendations for Future Reporting

No presentation of NAEP results for the press can be fully adequate. The degree of simplification that is needed to generate comprehensible NAEP reports, the lack of technical understanding on the part of press writers, and the pressure on press writers to reduce NAEP results to a small number of simple and eye-catching conclusions all guarantee some degree of oversimplification or misinterpretation by the press. However, the presentation of NAEP results in the articles reviewed here was clearly less than satisfactory, and the types of errors that appeared suggest a number of changes in reporting that might improve the interpretation of NAEP results in the future.

Provide Complements to Scaled Scores

The pervasive reliance on anchor points and achievement levels in the articles reviewed here confirms the importance of providing lay audiences with intuitively clearer metrics to complement scaled scores. Scaled scores alone do not allow lay audiences to make their own judgments about the adequacy of student performance or to understand the size and importance of differences over time or between groups.

Nonetheless, it is clear that the anchor-point and achievement-level methods, *as implemented and reported in the 1990 assessment*, were not fully adequate for this purpose: They provided useful metrics for writers, but they also engendered oversimplification and outright misinterpretations. Further exploration of alternative methods and of alternative presentations of the anchor points and achievement levels seems necessary.

Clarify the Difference Between Expected and Actual Performance

For somewhat different reasons, both the summer articles about anchor points and the autumn articles about achievement levels showed confusion between the actual and expected levels of performance. One clear source of these misinterpretations was vague language in the NCES report about the grade levels at which material is taught. These descriptions were internally inconsistent (referring both to the grades in which material is introduced and to the grades in which it is covered). More important, phrases such as "eighth-

grade material" have multiple meanings (e.g., that the material is taught in the eighth grade, that the typical eighth-grader has mastered the material, and that most eighth-graders have mastered the material). These multiple meanings were not clarified for readers, and readers were not told that the statements in the report were about curriculum rather than mastery. To avoid misinterpretations by the press, NAEP reports should clarify how, if at all, anchor points and achievement levels (or other points on the scale that are the focus of presentation) are linked to grade levels in terms of curriculum, expectations, or actual performance.

Find Clearer Ways of Presenting Actual Performance on Exemplar Items

The fundamental purpose of NAEP is to communicate what students know and can do. As Linn (1990) has pointed out, anchor points used alone generate serious misunderstandings—specifically, underestimates—of what higher-achieving students know and can do. Lay readers are likely to misconstrue the percentage of students reaching the anchor points as p-values for anchor items, even though, in the case of the higher anchor points, those p-values are typically far higher. Precisely the same error could occur with achievement levels.

Both NCES and NAGB presented p-values for illustrative items in 1991, but this review revealed that simply presenting p-values, even p-values for student groups defined by the levels or anchor points, was ineffective. Most of the press writers simply ignored the items altogether; of those who used the items, some presented no indication of success rates on them. In the autumn articles, most of the writers who presented estimates of success rates misconstrued the proportion of students reaching the achievement levels as p-values. Virtually none of the authors of the summer or autumn articles addressed the disparity between p-values and the percentages of students reaching the anchor points or achievement levels.

One source of this ineffectiveness may be that the NCES and NAGB reports largely left reporters to their own devices in disentangling the percentages of students reaching the anchor points or achievement levels, overall p-values, and p-values for subgroups defined by the levels. This is unrealistic: Most press writers lack both the expertise needed to understand the disparities and the time to explore them. The relationships among these measures must be explained simply and clearly in the NAEP reports themselves.

Highlight and Clarify the Continuity of Performance

Both the anchor-point and achievement-level methods encouraged writers to misrepresent performance as discontinuous. Clearly, if these methods are to remain a central aspect of reporting, the way in which they are presented needs to be improved. The scattered textual references to *consistent* performance in the NCES report simply did not suffice to solve this problem. Methods of clearly presenting and illustrating the continuity of performance—perhaps graphical methods in addition to verbal ones—need to be explored and made salient in official reports of the NAEP.

Improve the Reporting of Achievement Levels

In addition to the changes noted above, at least two important changes in the reporting of achievement levels appear necessary. First, the judgmental nature of the levels needs to be made clear, and the significance of this—in particular, that different judges or methods could lead to different levels—should be clarified explicitly.

Second, barring a sea change in reporting, the press will see the percentages of students reaching the levels as statements of what students know and can do. Therefore, accurate interpretation of the levels will require clear and defensible explanations of what students at the levels can do.

Consider the Styles of Presentation

Although no method of presentation is ideal, this review suggests several general guidelines for the presentation of NAEP results that might improve their interpretation by the press. Clearly, one cannot rely on fine points in the presentation of the results—such as the use of the word *should* in the descriptions of achievement levels or the reference to *consistent* performance in the discussion of anchor points—to have much influence on press reports. If points such as the consistency of performance are important, they should be made salient. Caveats in particular need to be made clear and salient.

Clarity and salience alone may be insufficient, however. One of the characteristics of many of the phrases in the NCES and NAGB reports that press writers picked up (or quoted outright) was that they were simple, clear, and easy to use with little or no modification. It may be prudent to provide simple and directly quotable statements of all the most important points, including caveats that NCES and NAGB hope writers will note.

The articles reviewed here also suggest the need to be cautious in using any novel or arcane metrics. In retrospect, it is not surprising that many press writers chose to use metrics such as grade equivalents in reporting NAEP, even when those metrics required some (presumably unintentional) distortion of the results presented. Grade equivalents (at least in the unspecific sense of "grade-level work") were familiar to the writers and to their readers, whereas anchor points and achievement levels were not.

When arcane metrics are desirable on other grounds, it will be necessary to take extra care, not only to make them clear to writers but to facilitate their use by writers. Some of the steps suggested above, such as clarifying p-values and the continuity of performance, might help in this respect.

Research Methods of Presentation

Finally, this review suggests the need for ongoing research on the presentation of NAEP results. The effectiveness of reporting methods must be established empirically. A variety of research methods, including focus groups, interviews, and reviews of the sort reported here, can be used from time to time to test the effectiveness of alternative methods of presentation and to fine-tune the methods deemed most appropriate.

Bibliography

Adams, C. (1991). Counting on success. *New Orleans Times Picayune*, October 7.

Asayesh, G. (1991). Report Card results are mixed for Maryland; math skills low, but test scores gain. *Baltimore Morning Sun*, October 1.

Baltimore Evening Sun (1991). Math study shows lots of problems; 8th graders have limited understanding, June 6.

Bourque, M. L., and Garrison, H. H. (1991). *The Levels of Mathematics Achievement: Initial Performance Standards for the 1990 NAEP Mathematics Assessment. Volume I: National and State Summaries.* Washington, D.C.: National Assessment Governing Board, September 30.

Burstein, L., Koretz, D. M., Linn, R. L., Sugrue, B., Novak, J., Lewis, E., and Baker, E. (1993). *The Validity of the 1992 NAEP Achievement Level Descriptions as Characterizations of Mathematics Performance.* Los Angeles, Calif.: UCLA, Center for the Study of Evaluation (CSE).

Cohen, M. (1991). Demand increases for national testing. Studies show gap between students today and goals. *Boston Globe*, October 6.

Cooper, K. (1991). U.S. youth fail math test. *Washington Post*, June 7.

DeWitt, K. (1991). Math scores in public schools shows no state is "Cutting It." *The New York Times*, June 7, p. 1.

Eskey, K. (1991). Numbers stymie most 8th graders. *Rocky Mountain News*, June 7.

Farmer, R. (1991). Virginia eighth-graders' scores better, but not up to standards. *Richmond Times-Dispatch*, October 1.

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

Hernandez, A. (1991). Math's a problem; tests show "We compare favorably," but educators aren't happy. *Rocky Mountain News*, June 7.

Hildebrand, J. (1991). U.S. study: Math scores still aren't adding up. *Newsday*, September 30.

Jaeger, R. M. (1992). General issues in reporting the results of the NAEP Trial State Assessment results. In National Academy of Education, *Assessing Student Achievement in the States: Background Studies*. Stanford, Calif.: Stanford University, National Academy of Education.

Koretz, D. M. (1989). NAEP scales: How useful are they? Paper presented at the annual assessment conference of the Education Commission of the States, Boulder, Colo., June.

Koretz, D. M., and Lewis, E. (unpublished). Indicators of achievement in mathematics and science. Santa Monica, Calif.: RAND, unpublished research.

Linn, R. L. (1990). Historical origins and issues in the National Assessment of Educational Progress. In *Assessment at the National Level*, symposium presented at the Institute for Practice and Research in Education, University of Pittsburgh, Penn., October 26.

Linn, R. L., Koretz, D. M., Baker, E. L., and Burstein, L. (1992). *The Validity and Credibility of the Achievement Levels for the 1990 National Assessment of Educational Progress in Mathematics*. Los Angeles, Calif.: UCLA, Center for the Study of Evaluation, CSE Technical Report 330.

Lue, A. (1991). State math skills don't measure up. *Sacramento Bee*, June 7.

Mullis, I. V. S., Dossey, J. A., Foertsch, M.A., Jones, L. R., and Gentile, C. A. (1991a). *Trends in Academic Progress*. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics, Report No. 21-T-01, November.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G. W. (1991b). *The State of Mathematics Achievement*. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics, Report No. 21-ST-04, June.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., and Phillips, G. W. (1993). *NAEP 1992 Mathematics Report Card for the Nation and the States*. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics, Report No. 23-ST02, April.

National Education Goals Panel (1991). *The National Education Goals Report, 1991: Building a Nation of Learners*. Washington, D.C.

Nelis, K. (1991). State 8th graders don't measure up in math. *Times Union*, June 7.

O'Brien, K. (1991). Teacher training, test scores tied to students' math scores prompt look at teachers' credentials. *Charlotte Observer*, October 2.

Orlando Sentinel (1991). Nation's Report Card comes under fire. October 2.

Phillips, G. W., Mullis, I. V. S., Bourque, M. L., Williams, P. L., Hambleton, R. K., Owen, E. H., and Barton, P. E. (1993). *Interpreting NAEP Scales*. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics, April.

Ribadeneira, D. (1991). State education officials say no to national test. *Boston Globe*, September 5.

Richmond News Leader (1991). Mediocre proficiency reported in math, reading, writing, science. September 30.

Roser, M. A. (1991a). Math skills ring alarm bell. Scores equal problems: N.C. 8th-graders rank 36th of 40. *Charlotte Observer*, June 7.

Roser, M. A. (1991b). Report charts U.S. students' shortcomings in math. *Philadelphia Inquirer*, June 7.

Roser, M. A. (1991c). Tests results add up to this: U.S. students lagging in math. *Miami Herald*, June 7.

Roser, M. A. (1991d). U.S. kids falling behind in math. *San Jose Mercury News*, June 7.

San Jose Mercury News (1991). American students are 20 years behind. Study says they don't handle basics well. September 30.

Seattle Times (1991). Looking for the answers. Math: A nation of dunces. July 14.

St. Paul Pioneer Press Dispatch (1991). Researcher says politics tainted report. October 2.

Stufflebeam, D. L., Jaeger, R. M., and Scriven, M. (1991). *Summative Evaluation of the National Assessment Governing Board's Inaugural 1990–91 Effort to Set Achievement Levels on the National Assessment of Educational Progress*. Kalamazoo, Mich.: Western Michigan University Evaluation Center.

U.S. General Accounting Office (1992). *National Assessment Technical Quality.* Washington, D.C., GAO/PEMD-92-22R, March 11.

USA Today (1991). 8th grade math scores: State by state comparison. June 7.

White, B. (1991). U.S. students improving in science, math, but worse in language skills. *Atlanta Journal*, September 30.