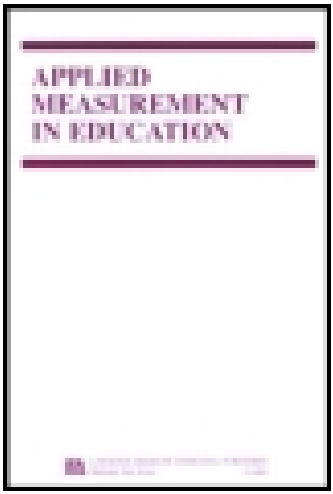


Applied Measurement in Education

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hame20

Student Test Score Reports and Interpretive Guides: Review of Current Practices and Suggestions for Future Research

Dean P. Goodman & Ronald K. Hambleton

Published online: 07 Jun 2010.

To cite this article: Dean P. Goodman & Ronald K. Hambleton (2004) Student Test Score Reports and Interpretive Guides: Review of Current Practices and Suggestions for Future Research, Applied Measurement in Education, 17:2, 145-220, DOI: 10.1207/s15324818ame1702_3

To link to this article: http://dx.doi.org/10.1207/s15324818ame1702_3

APPLIED MEASUREMENT IN EDUCATION, *17*(2), 145–220

Student Test Score Reports and Interpretive Guides: Review of Current Practices and Suggestions for Future Research

Dean P. Goodman and Ronald K. Hambleton
Center for Educational Assessment
University of Massachusetts Amherst

A critical, but often neglected, component of any large-scale assessment program is the reporting of test results. In the past decade, a body of evidence has been compiled that raises concerns over the ways in which these results are reported to and understood by their intended audiences. In this study, current approaches for reporting student-level results on large-scale assessments were investigated. Recent student test score reports and interpretive guides from 11 states, three U.S. commercial testing companies, and two Canadian provinces were reviewed. On the basis of past score-reporting research, testing standards, and the requirements of the *No Child Left Behind Act of 2001,* a number of promising and potentially problematic features of these reports and guides are identified, and recommendations are offered to help enhance future score-reporting designs and to inform future research in this important area.

Large-scale assessments have played a prominent role for many years in America's kindergarten to Grade 12 school systems (Hamilton & Koretz, 2002; Linn, 1998), informing a wide-range of national, state, and local reform efforts designed to improve student learning. Over this time, a great amount of attention has been directed toward the creation of technically sound assessments that can stand up to intense public and professional scrutiny. Considerably less attention, however, has been given to ways in which the results of the assessments are organized, reported,

Requests for reprints should be sent to Dean P. Goodman, Center for Educational Assessment, 152 Hills South, University of Massachusetts, Amherst, MA 01003. E-mail: dgoodman@psych.umass.edu

and used (Hambleton, 2002b). Concerns about the reporting of assessment results have been raised in recent years because there is a body of evidence that shows confusion among policymakers, educators, and the public over the meaning and interpretation of large-scale assessment results (Hambleton, 2002b; see also Hambleton & Slater, 1997; Jaeger, 1998).

In the next several years, states will be reporting assessment results to a larger and more diverse audience than ever before. To comply with the *No Child Left Behind Act of 2001* (NCLB), states must report results on mathematics, reading, and science assessments at the state, district, school, subgroup, and individual student levels across a wide range of grades. By the 2005–2006 school year, assessment reports will be distributed annually to parents, guardians, and teachers of an estimated 22 million students in Grades 3 to 8 alone (Landgraf, 2001). This widespread distribution of assessment results—and the expectation that they will play a critical role in ensuring that students obtain the knowledge, skills, and abilities expected of students in their respective grades—will lead to unprecedented amounts of attention being directed toward state assessment results, especially at the individual student level.

Very little research currently exists on how student-level results from large-scale kindergarten to Grade 12 assessments are reported. Given the increased role these results will play in the United States as a consequence of NCLB and the available evidence that shows the difficulties that many people have in understanding large-scale assessment results, there is a clear need to identify effective ways to report student-level results on large-scale assessments.

In this article, we review student score reports and related interpretive guides from a sample of states, Canadian provinces, and U.S. commercial test publishers. The purposes of this review are (a) to determine the types of information currently included in student score reports and interpretive guides and to describe the ways this information is presented, (b) to identify promising and potentially problematic features of these reports and guides, and (c) to offer recommendations that may enhance future reporting practices.

CURRENT REQUIREMENTS AND GUIDELINES RELEVANT TO STUDENT-LEVEL REPORTING ON STATEWIDE TESTS

To help inform the review of recent student-level score reports and to highlight resources that will assist states in their reporting efforts, key legislative and professional requirements and research-based guidelines relevant to student-level score reports will be considered first.

The Legislative Requirement to Report Individual Student Results on Statewide Assessments: Conditions of the *No Child Left Behind Act of 2001*

In recent years, the reporting of student-level results on statewide assessments has become widespread. A review of 50 state profiles compiled by Goertz, Duffy, and Carlson-LeFloch (2001) showed that 45 states report student-level results on one or more statewide tests. Many states attach significant stakes to these results. In 2004, 20 states make graduation contingent upon performance on statewide tests (Education Week, 2003). In seven states, students must pass a statewide test to be promoted to the the next grade. The release of student-level results and the high stakes that often accompany them have undoubtedly helped raise the profile of state assessments among educators, parents, students, and the general public. With the signing into law of NCLB in January 2002, the amount of attention that will be directed toward state assessments and individual student results is likely to increase even further over the next few years.

Regarded as the most significant federal education policy initiative in a generation (Illinois State Board of Education, 2002), NCLB outlines a wide range of goals to ensure that each child in the United States is able to meet the high learning standards of the state in which he or she lives. Accountability is the centerpiece of NCLB, with statewide assessments playing a critical role in ensuring that the school system is accountable for the performance of all students. Under this law, states are required to administer high-quality annual assessments in mathematics and reading or language arts to all students in Grades 3 through 8 by the 2005–2006 school year, extending the existing requirement that students be tested in these subject areas at least once during Grades 3 through 5, Grades 6 through 9, and Grades 10 through 12 (NCLB, 2001, §1111[b][3][C][v][I] and §1111[b][3][C][vii]). Beginning in 2007–2008, states also will be required to measure the proficiency of each student in science at least once during Grades 3 through 5, Grades 6 through 9, and Grades 10 through 12 (NCLB, 2001, §1111[b][3][C][v][II]).

Under NCLB, individual results must be reported for all students who take part in the annual assessments (estimated by Landgraf, 2001, to be a staggering 22 million students in Grades 3 through 8 alone). Specifically, states are required to:

> produce individual student interpretive, descriptive, and diagnostic reports…that allow parents, teachers, and principals to understand and address the specific academic needs of students, and include information regarding achievement on academic assessments aligned with State academic achievement standards, and that are provided to parents, teachers, and principals, as soon as is practicably possible after the assessment is given, in an understandable and uniform format, and to the extent practicable, in a language that parents can understand. (NCLB, 2001, §1111[b][3][C][xii])

These reports must "describe student achievement measured against the state's academic achievement standards" (Title I–Improving the Academic Achievement of the Disadvantaged Final Rule, 2002, p. 45038) and should "be consistent with relevant, nationally recognized professional and technical standards" (NCLB, 2001, §1111[b][3][C][iii]).

Professional and Technical Standards Relevant to Reporting Individual Student Results on Statewide Assessments

At least three resources are pertinent to the NCLB requirement that state assessments and individual student reports shall be consistent with relevant professional and technical standards. These include *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999), *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices [JCTP], 2004), and *Code of Professional Responsibilities in Educational Measurement* (NCME, 1995).

The primary source of professional and technical standards that guide most aspects of testing is the *Standards for Educational and Psychological Testing* (AERA et al., 1999). The purpose of this resource is to "provide criteria for the evaluation of tests, testing practices, and the effects of test use" and to offer a frame of reference that, in concert with professional judgment, can be used "to assure that relevant issues are addressed" (p. 2). A significant number of standards outlined in *Standards for Educational and Psychological Testing* are relevant to reporting individual student results on large-scale assessments. For reference purposes, a list of the 10 particularly relevant standards is provided in Appendix A.

The *Code of Fair Testing Practices in Education* (JCTP, 2004) is a recent publication that represents the spirit of selected portions of the *Standards for Educational and Psychological Testing* (AERA et al., 1999) in a manner that is relevant and meaningful to test takers and their parents or guardians. This document replaces the first edition of the *Code,* which was published in 1988 (JCTP, 2004). A list of 15 principles relevant to reporting and interpreting assessment results and informing test takers about important aspects of the assessment is provided in Appendix B.

A document that provided a basis for elements of the *Standards for Educational and Psychological Testing* is the *Code of Professional Responsibilities in Educational Measurement* (NCME, 1995). Appendix C includes a list of 11 professional responsibilities of those who interpret, use, and communicate assessment results.

Literature Related to Score Reporting

Complementing the legislative requirements and professional and technical standards is a growing body of literature on score reporting and the effective display of quantitative information. Hambleton (2002b); Hambleton and Slater (1997); Impara, Divine, Bruce, Liverman, and Gay (1991); Jaeger (1998); the National Education Goals Panel (NEGP, 1998); the National Research Council (NRC, 2001); and Wainer, Hambleton, and Meara (1999) provided clear evidence that many users of assessment data have difficulty interpreting and understanding results presented in large-scale assessment reports. Although most current research on score reporting is based on reports from the National Assessment of Educational Progress (NAEP), many findings and principles that have emerged from this research are relevant to student-level score reports and will be summarized next.

Findings From Research on National Assessment Reports

Considerable interest in researching the accessibility and comprehensibility of public reports for NAEP has been shown in recent years. Although these reports have improved since approximately 1992, some early concerns are worth reviewing.

A primary problem of NAEP reports from the early 1990s was that they assumed "an inappropriately high level of statistical knowledge for even well-educated lay audiences" (NRC, 2001, p. 88). Too many technical terms, symbols, and concepts were required to understand the message underlying even simple data (NRC, 2001, p. 88). As observed by Hambleton and Slater (1997), statistical jargon (e.g., statistical significance, variance, standard error) confused and even intimidated some users. Symbols (e.g., "<" and ">" to denote statistically significant differences) and technical footnotes were misunderstood or ignored by many users of the reports (Hambleton & Slater, 1997).

Other major criticisms of the NAEP reports included presenting "too much information, making it difficult for readers to find and extract what they really want to know" (NRC, 2001, p. 89) and including "overly dense displays that readers find daunting" (p. 89). Past reports were criticized for not making "enough use of graphical alternatives to textual and tabular formats" (p. 90). Even when attempts were made to redesign the displays for easy access (e.g., using three-dimensional bar and pie charts), they sometimes led to such problems as increased clutter or perceptual inaccuracies (p. 89). Other concerns included the lack of descriptive information (e.g., definitions and concrete examples) that would have helped provide meaning to the assessment results (Hambleton & Slater, 1997).

General Principles for Effectively Reporting Large-Scale Assessment Results

A number of general principles for effectively reporting large-scale assessment results can be gleaned from recent score-reporting literature (e.g., Hambleton, 2002b; Hambleton & Slater, 1997; Jaeger, 1998; NRC, 2001; Snodgrass & Salzman, 2002; Wainer, 1997a; Wainer et al., 1999; Ysseldyke & Nelson, 2002) and literature relating to the visual display of quantitative information (e.g., Tufte, 1983, 1990; Tukey, 1990; Wainer, 1990, 1992, 1997b; Wainer & Thissen, 1981). These principles include (a) making the report readable, concise, and visually attractive; (b) keeping the presentation clear, simple, and uncluttered; (c) not trying to do too much with a data display (i.e., displays should be designed to satisfy a small number of preestablished purposes); (d) including text to support and improve the interpretation of charts and tables; (e) minimizing the use of statistical jargon; (f) including a glossary of key terms; (g) using bar charts to facilitate comparisons; (h) grouping data in meaningful ways; (i) using boxes or graphics to highlight main findings; (j) avoiding the use of decimals; (k) using color in a purposeful manner (given the potential for misuse, however, the general use of color was not universally recommended); (l) piloting the reports with members of the intended audience; and (m) creating specially designed reports for different audiences.

Findings and Recommendations From the Literature on Student Score Reports

In a resource written when standards-based assessments and student-level reporting were not as widespread as they are today, NEGP (1998) outlined ways states could better inform parents about issues related to standards and state assessments and how states could report statewide and individual student results in more meaningful ways.

NEGP (1998) argued that "too often…individual student reports are not very clear" (p. 35) and are guilty of providing either too little information (e.g., a score or classification without any explanation of what the score or classification meant) or too much information (e.g., excessive details that made it difficult for parents to understand how their child performed). To achieve an appropriate balance, NEGP recommended that states answer four questions on their student reports:

1. How did my child do?
2. What types of skills or knowledge does his or her performance reflect?
3. How did my child perform in comparison to other students in the school, district, state, and, if available, the nation?
4. What can I do to help my child improve? (p. 36)

To help answer these questions and to provide additional contextual information with the test results (such as the purpose of the test, definitions of achievement levels, scoring guides, and what the test looked like), NEGP (p. 38) suggested states include an interpretive guide with the individual score reports. Emphasizing that interpretive guides should not take the place of informing parents *before* the administration of the assessment, NEGP (p. 38) recommended these guides as a way to provide parents with important information that would not likely fit on a single page (e.g., the reverse side of a score report).

NEGP (1998) outlined a number of other ideas that would help states report individual student results in a meaningful manner. These included encouraging parents to contact their child's teacher for more information about the child's test results, encouraging parents to ask questions about the school's educational practices (e.g., by including questions parents might ask on the student report or accompanying interpretive material), emphasizing the importance of looking at a variety of sources of information when evaluating student performance, and providing examples of student work and test questions that illustrate what students know and should be able to do.

Comments from a small focus group composed of 11 parents from across the United States also were reported by NEGP (1998). As part of this focus group study, parents were asked to review and comment on six individual student reports produced by commercial test publishers. Although the small sample size limits the extent to which the findings can be generalized, comments on what parents liked and disliked about the reports are worth noting. In general, parents involved in the study appreciated explanations of what the scores on the test meant and liked being able to tell at a glance how their child performed. They also liked seeing subtest scores and descriptions of the skills that were assessed by the test. Parents appreciated learning what could be done to improve a student's score. They did not like reports that were too technical (e.g., containing statistical jargon and complex definitions) or reports that did not give recommendations on what they should do with the test results. They also raised concerns about small fonts that made parts of the reports difficult to read.

Impara et al. (1991) investigated the extent to which teachers in one state were able to interpret student-level results on a standardized state assessment and the extent to which interpretive information provided on the reverse side of the student score report helped to improve teacher understanding. Although many teachers provided reasonable interpretations of information contained on the score reports, some types of information were misunderstood by large numbers of teachers. As noted by Impara et al., "areas of weakness related to scale and normal curve equivalent (NCE) scores; the percentile band performance profile; interpreting grade-equivalent scores; and the norm-group number correct on the skills chart, which provides the average number correct by the national norm group and the number correct by the student" (p. 17). Regardless of the availability of interpretive infor-

mation, most teachers (75% with interpretive information and 82.5% without such information) could not properly interpret percentile band performance profiles. Impara et al. noted that interpretive information helped address many, but not all, of the difficulties teachers had in interpreting the other scores.

Impara et al. (1991) showed that interpretive material helped facilitate teachers' understanding of student scores on a standardized state assessment. Still, problems remained even when interpretive material was provided (e.g., even with interpretive material, teachers did not understand the meaning of percentile bands that overlapped). In addition to recommending more research on teachers' understanding of student score reports, Impara et al. suggested that some problems in score interpretation might disappear if score reports contained only instructionally relevant information (e.g., they recommended removing rarely used scores such as the NCE to make the reports less intimidating for teachers and parents).

A recent publication by Forte Fast and the Accountability Systems and Reporting State Collaborative on Assessment and Student Standards (2002) should be especially helpful to states. This resource, sponsored by the Council of Chief State School Officers, was designed to help states and local agencies meet the reporting requirements of NCLB and to help them design public reports that effectively communicate accountability, assessment, and other educational indicators in an easily understood manner. It provides some excellent guidelines and illustrations that can help state and local agencies improve their reporting practices.

Current Student Score Reports and Interpretive Guides: Methods and Results

Given the concerns raised in the score-reporting literature, the lack of information on how states are reporting individual student results on large-scale assessments, and the expected increase in the number of student-level reports that will be produced and distributed in the next several years, a review of the ways student-level assessment results are reported appeared warranted. Recommendations that might follow from this review may be influential in the design of new reports.

METHODS

Data Collection

Student reports and accompanying interpretive material were requested from the departments of education from 14 states, three U.S. commercial testing companies, and the departments of education from two Canadian provinces. Reports and interpretive material for Grade 10, or the grade closest to Grade 10, were requested. Grade 10 was chosen because it is a grade in which student-level results

on large-scale assessments are commonly reported, often with considerable stakes attached (e.g., high school graduation).

States were selected to represent a cross-section of states from across the country with low, medium, and high populations. Responses were received from 11 of the 14 states. The participating states were Connecticut, Delaware, Louisiana, Massachusetts, Minnesota, Missouri, New Jersey, Pennsylvania, Virginia, Wisconsin, and Wyoming. Material also was obtained from three publishers of widely administered commercial tests. These were Harcourt Educational Measurement (publisher of the *Stanford 10*), CTB/McGraw-Hill (publisher of the *TerraNova, The Second Edition*), and Riverside Publishing (publisher of the *Iowa Tests of Educational Development* [*ITED*]). British Columbia and Ontario, two of seven Canadian provinces that report student-level results on provincewide assessments (Taylor & Tubianosa, 2001), also submitted material for the study. It was hypothesized that reviewing material from departments of education outside of the United States would offer additional insights into different ways of reporting student-level results on large-scale assessments.

Two types of student-level reports (a home report and a more detailed student profile report) from the commercial testing programs were used in the study (Harcourt Educational Measurement, 2002, p. a–b; CTB/McGraw-Hill, 2001a, p. 20, 23; University of Iowa, 2001a, 2001b, p. 40); complete interpretive material was received for the home reports only (Harcourt Educational Measurement, 2002, p. a; CTB/McGraw-Hill, 2001b, 2003; University of Iowa, 2001b). For each state and province, single reports and all accompanying interpretive guides written for parents and guardians were used in the study (British Columbia Ministry of Education, 2002; Connecticut State Board of Education, 2002a, p. 12–13, 2002b; CTB/McGraw-Hill, 1997; Data Recognition Corporation, 2003a, 2003b; Delaware Department of Education, 2002; Education Quality and Accountability Office, 2002a, 2002b; Louisiana Department of Education, 2002a, 2002b; Massachusetts Department of Education, 2002a, 2002b; Minnesota Department of Children, Families & Learning, 2002; New Jersey Department of Education, 2002; Pennsylvania Department of Education, 2002; Virginia Department of Education, 2002; Wisconsin Department of Public Instruction, 2002, p. 4–5; Wyoming Department of Education, 2002a, p. 23–24, 2002b).

Data Analysis

An iterative content analysis procedure was used to analyze and summarize the student reports and the accompanying interpretive guides. The reports and interpretive guides were first reviewed and analyzed individually. Key features of each report and interpretive guide were identified, including those relevant to NCLB requirements, professional standards, and guidelines from previous score-reporting research.

After all reports and interpretive guides were analyzed individually, a category coding system was created that addressed key features across these score reports and guides. The student reports and accompanying interpretive materials were reviewed again and analyzed in terms of four coding categories:

1. Contextual information, which included (a) grade of report and year of distribution, (b) stakes attached to student results, and (c) overlap between reports of states and commercial test publishers;
2. General design features, which included (a) the physical characteristics of the student reports, (b) the availability and physical characteristics of any accompanying interpretive guides, and (c) methods used to organize the material;
3. Types of information included in the student reports and the manner in which they were reported, which included (a) the number of subjects reported, (b) students' overall scores, (c) overall results in relation to performance levels, (d) diagnostic information (operationally defined as information that was more detailed than overall results for general subject areas), (e) comparative information, and (f) information regarding the precision of the test results; and
4. Types of information included in the interpretive material and the manner in which they were reported, which included (a) answers to questions parents might have about the assessment (What was the purpose of the test? What was assessed? What did the test look like? Where can parents get more information? What can parents and guardians do to help students improve?) and (b) details regarding key elements of the student reports (descriptions of component parts of the reports and definitions of technical terms).

Validation of the Results

Once all of the materials were analyzed, a preliminary report that outlined the findings and implications of the study was distributed to representatives of each participating state, province, and commercial test publisher for critical review. Representatives from six states, one province, and all three commercial test publishers submitted feedback about the report. Based on this feedback, minor revisions were made to the report to help clarify the findings and implications of the study.

RESULTS

The results of this study are organized around the four coding categories defined earlier. Consistent with the terms of participation in the study, the names of states, Canadian provinces, and commercial test publishers were not identified when their reports were reviewed and discussed. Where promised, the names of the states, Ca-

nadian provinces, and commercial test publishers and any other obvious identifiers were removed from illustrative examples.

Contextual Features of Reports and Interpretive Guides

Grades and Year of Release

Four of the 11 states submitted 10th-grade student score reports and accompanying interpretive material. The remaining states either submitted 11th-grade reports (four states) or generic student reports from lower grades that were comparable to the reports produced for the 10th grade (three states). The commercial test publishers provided generic sample reports that were used across a range of grades (including Grade 10). Both Canadian provinces submitted 10th-grade student reports. With the exception of two commercial test reports published in 2001, all student reports included in this study were released in 2002.

Stakes Assigned to Results

The stakes assigned by states and Canadian provinces to student-level results were identified either from the submitted material or the Web sites of the relevant departments of education. Low or no stakes were attached to student-level results in four states. High stakes (operationally defined as either a requirement for promotion to the next grade or graduation) were attached to assessment results in six states. In one state, assessment results were included on students' transcripts but without any explicit state-sanctioned stakes.

No stakes were attached to students' individual assessment results in one of the Canadian provinces, although high stakes (graduation from high school) were attached to the results in the other Canadian province. No explicit stakes were assigned by the commercial test publishers to the individual student results because stakes are determined by states and local districts, not by the test publishers.

Overlap Between Reports of States and Commercial Test Publishers

Some overlap in design and content was noted across the reports of some states and commercial test publishers. Reports from two states were published by one of the commercial test publishers. These reports shared some common features with the commercial test publisher's own reports and each other (e.g., layout and the provision of student and test administration information) but were distinct in a number of important ways relating to the types of information included in the reports and the manner in which results were reported.

Four states reported results from assessments developed by two of the commercial test publishers. Some results reported by three states (e.g., national percentile

results on the multiple-choice components of the tests) were based on the assessments of one of these two test publishers. The results reported by one state were based entirely on one of the commercial tests.

General Design Features

Physical Characteristics of the Student-Level Reports

The student-level reports from nine states consisted of two letter-sized (8½″ × 11″) pages (which could be distributed on one double-sided sheet of paper). The reports from two states consisted of one letter-sized page.

The student-level report from one Canadian province was two letter-sized pages (distributed on one double-sided sheet of paper). The report from the second Canadian province was four letter-sized pages (which could be distributed as a double-sided 11″ × 17″ pamphlet folded in half).

Reports from two commercial test publishers consisted of two-letter sized pages (distributed on one double-sided sheet of paper). One of the reports from the third commercial test publisher was one letter-sized page and the other report was four letter-sized pages (distributed on a double-sided 11″ × 17″ pamphlet folded in half).

Two commercial test publishers produced very colorful score reports. The score report of the third commercial test publisher contained some color, but this color was limited to the title page and headings of the report (color was not used in the presentation of assessment results). States used little or no color in their student score reports (fewer than one half of the states used more than one color in their score reports). The two Canadian provinces produced black-and-white score reports.

Availability and Physical Characteristics of Interpretive Guides

Some form of interpretive material accompanied the student-level reports of all states, Canadian provinces, and commercial test publishers. Seven of the 11 states, both Canadian provinces, and two of the three commercial test publishers included interpretive material on the actual student-level report (typically one page of material on the back of the report, although a report from one of the commercial test publishers included approximately four pages of interpretive material). Five states, one province, and one commercial test publisher produced a separate interpretive guide. One state, one province, and one commercial test publisher produced two complementary interpretive guides for parents and guardians. One state supplemented information provided on the back of the student report with an interactive Web-based interpretive guide. A similar strategy was used by one commercial test

publisher, which supplemented its separate printed guide with an interactive Web-based version. One Canadian province supplemented interpretive material included in the student report (e.g., a glossary of key terms) with a separate interpretive guide. One commercial test publisher that included interpretive material on the back of its report indicated it plans to release a separate interpretive guide in the near future.

The separate interpretive guides produced by four states were 4, 14, 20, and 37 pages long. The four-page guide was printed on an 11″ × 17″ pamphlet folded in half (allowing the student report to be inserted inside the guide). The 14-page guide was 5½″ wide and 8½″ tall. All other guides produced by states were letter-sized.

The separate interpretive guide for one of the Canadian provinces was seven letter-sized pages. The separate guide produced by the commercial test publisher was 12″ wide and approximately 19″ tall, which was folded in half to create a folder for the student report (this folder also included an inside pocket and resealable flap that held the report in place). This commercial test publisher also produced a four-page guide for one of its student-level reports, but the complete guide was not included in the material submitted for the study.

The interpretive guides of two commercial test publishers were in color (although only one made significant use of color). With a few exceptions, states and Canadian provinces did not use color in their interpretive guides.

Methods Used to Organize the Material

States, Canadian provinces, and commercial test publishers used a variety of techniques to separate different components of the reports and interpretation guides. The most common technique was the use of descriptive headings, which were used to some extent on all documents (one state made minimal use of headings in its interpretive guide). Typically, these headings were phrased as simple statements (e.g., Overall Results, Results by Academic Standards); two states and one Canadian province used questions parents might ask as headings on their score reports (e.g., How did *student's first name inserted here* do on this test? What are your child's strengths and weaknesses?). This is a concept that Wainer pioneered—he viewed personalizing the report as important. Sections in all reports also were separated by boxes, lines, or dark bars with white headings (see Figure 1 for an example of how boxes were used to organize the results from one commercial score report). Several states and all commercial test publishers used color to help separate different components of the score reports or interpretive guides. White space helped minimize clutter in many documents; however, a lack of white space made three state reports and four state interpretive guides appear quite dense. A table of contents was included in one state's interpretive guide but was not included in other documents.

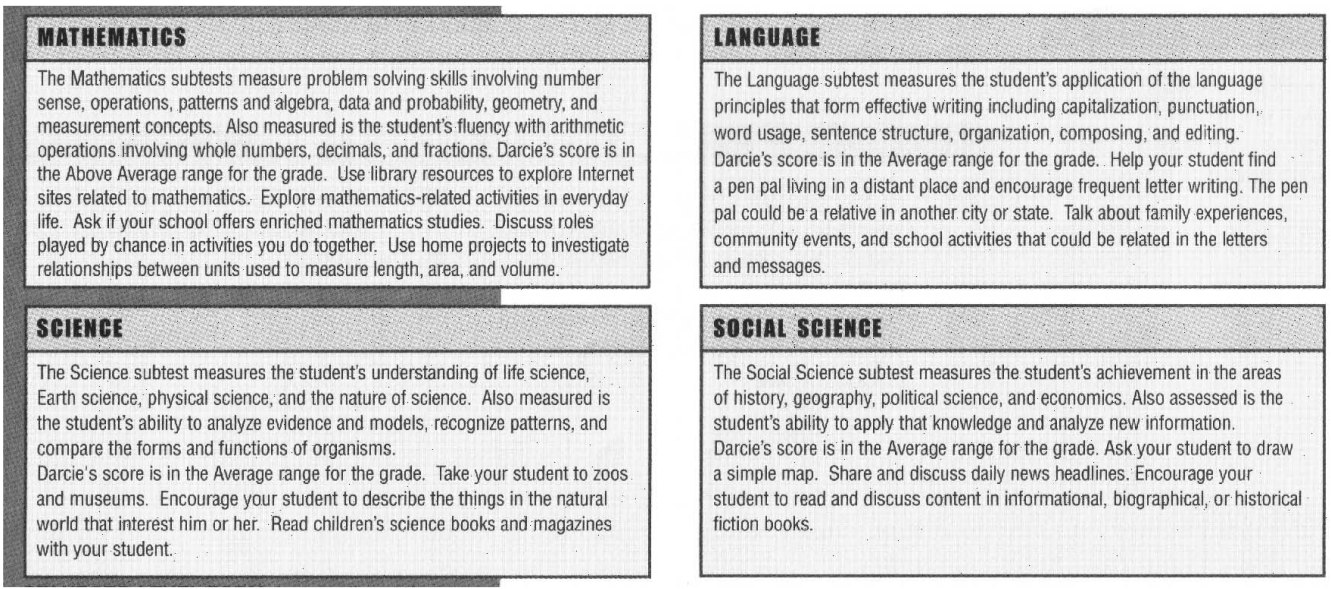


FIGURE 1 Reporting results using narrative text and boxes (display from a commercial test publisher's score report). Reproduced from the *Stanford Achievement Test: Tenth Edition.* Copyright © 2003 by Harcourt Educational Measurement, a Harcourt Assessment Company. Reproduced by permission. All rights reserved.

Types of Information Reported in Student Score Reports and Methods of Reporting

Number of Subjects

Student reports from 8 of the 11 states included results for multiple-subject areas (e.g., mathematics, reading, writing, science, social studies), ranging from two subject areas in three states to five subject areas in 2 states. Three states reported results for a single subject area (English language arts, mathematics, or science).

The reports from the two Canadian provinces included results for multiple subject areas (reading, writing, and numeracy in one Canadian province and reading and writing in the other province). Student-level reports from the commercial testing programs contained results for a larger number of subject areas than reports from any state or Canadian province (six subjects on two commercial reports and eight subjects on the third commercial report). All three commercial tests also reported composite scores, which comprise results of multiple subject areas.

Reporting Overall Scores on Student-Level Reports

Overall scores for a subject area were reported in the student score reports of all 11 states and the 3 commercial test publishers. For one province, no overall scores were reported (results were reported only in relation to three performance standards). For the other province, overall scores were reported only for those students who did not pass the relevant component of the assessment (all other students received a statement that indicated they passed the relevant component of the assessment).

Types of overall scores reported. Many different types of overall scores were used across the state, provincial, and commercial test reports reviewed in this study. Commonly reported overall scores included scaled scores, raw scores, number correct scores, percent correct scores, holistic scores (for writing), percentile rank scores, and stanines.

In four states, more than one type of overall score was provided on the student-level score reports. These included two states that reported two types of overall scores (scaled and percentile rank scores or scaled and number correct scores) and two states that reported three different types of overall scores (scaled score, national stanine, and national percentile or scaled score, national percentile, and raw score). Six states reported only overall scaled scores, and one state and one Canadian province reported only overall raw scores (total points achieved across all items). Across states, the most commonly reported overall scores were scaled scores, with all but one of the states reporting this type of score.

For two commercial test publishers, the number of overall scores reported varied across their home reports and more detailed student profile reports. The student profile reports for these publishers contained either 4 or 10 types of overall scores

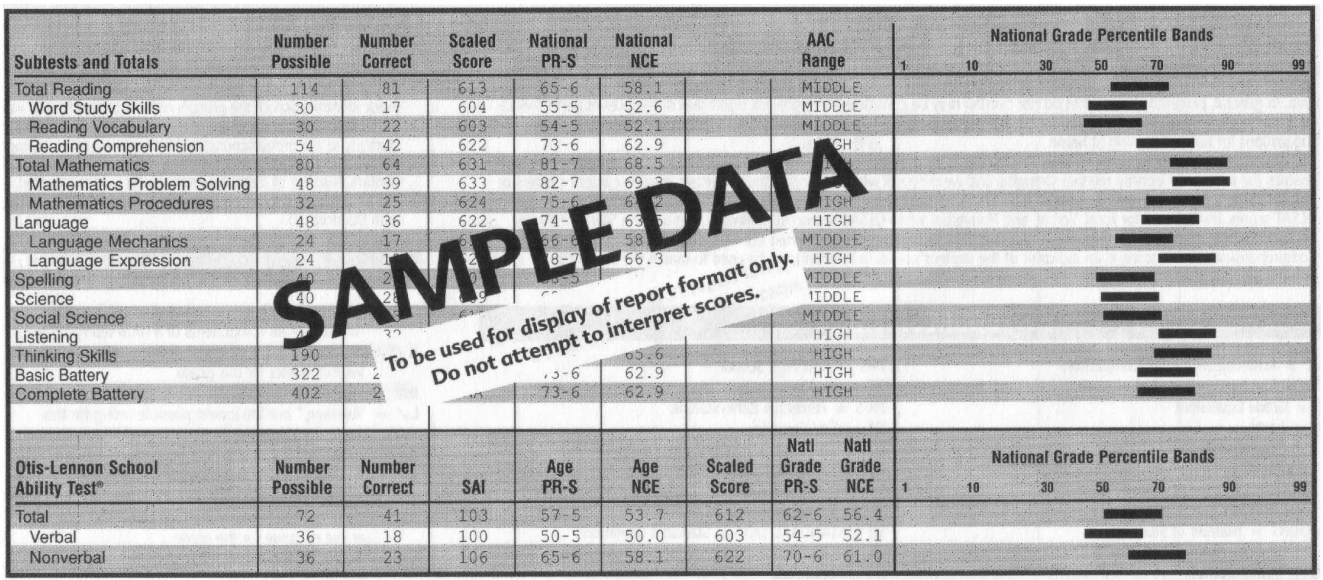
(e.g., scale scores, grade equivalent scores, national and age stanines and percentiles, number correct, normal curve equivalent, Lexile measures, school ability index). An example display from one of these reports is shown in Figure 2. In contrast, the home reports for these test publishers contained only 1 (national percentile) or 2 (national percentile and Lexile measure) overall scores. An equivalent display from one of these simpler versions is shown in Figure 3. Both score reports from the other commercial test publisher included only percentiles.

Methods of reporting overall scores. Overall scores were typically reported in both a numerical and graphical manner and in some cases were embedded within a narrative description of the student's performance. Ten of the 11 states and all 3 commercial test publishers reported overall scores in multiple ways (however, one of the commercial test publishers reported only numeric scores in one of the two score reports reviewed in this study—see Figure 4). One state reported only overall scores numerically. The one Canadian province that reported overall scores embedded these results within a short sentence (e.g., Your total reading score is ____ points.). Figures 5 and 6 illustrate other ways overall scores were reported by a state and a commercial test publisher.

Information regarding the precision of overall scores. Four of the 11 states and two of the three commercial test publishers provided information about the precision of at least one type of overall test score; the one Canadian province that reported overall scores did not provide information about its precision. This information was reported graphically by 2 states and one commercial test publisher, numerically by 1 state, and both numerically and graphically by 1 state and one commercial test publisher. An example of how 1 state reported the precision of overall scores (by including the standard error associated with a student's scaled score) is displayed in Figure 5. Figures 7 and 8 illustrate other ways this information was reported.

Strategies to provide meaning to overall scores. Whenever overall scores were reported, one or more strategies were used to help provide meaning to these scores. The most common strategy used by states and Canadian provinces was to report overall scores in relation to performance levels (described in the next section) and to describe the skills and knowledge that each performance level represented. Another popular strategy was to report a student's overall score in relation to scores of relevant comparison groups (e.g., average scores of students in the school, district, and state, see Figure 9); this strategy was used by nine states and all three testing companies but was not used by either Canadian province.

Other strategies used to provide meaning to overall scores included describing skills and knowledge measured by the test (e.g., see Figure 1); describing skills and knowledge typically possessed by students who obtained a particular overall score



Subtests and Totals	Number Possible	Number Correct	Scaled Score	National PR-S	National NCE	AAC Range	National Grade Percentile Bands
Total Reading	114	81	613	65-6	58.1	MIDDLE	
Word Study Skills	30	17	604	55-5	52.6	MIDDLE	
Reading Vocabulary	30	22	603	54-5	52.1	MIDDLE	
Reading Comprehension	54	42	622	73-6	62.9	HIGH	
Total Mathematics	80	64	631	81-7	68.5	[illegible]	
Mathematics Problem Solving	48	39	633	82-7	69.3	[illegible]	
Mathematics Procedures	32	25	624	75-6	61.2	HIGH	
Language	48	36	622	74-6	63.5	HIGH	
Language Mechanics	24	17	61[illegible]	66-6	58.[illegible]	MIDDLE	
Language Expression	24	19	62[illegible]	78-7	66.3	HIGH	
Spelling	40	2[illegible]	60[illegible]	5[illegible]	[illegible]	MIDDLE	
Science	40	28	609	6[illegible]	[illegible]	MIDDLE	
Social Science	[illegible]	[illegible]	61[illegible]	[illegible]	[illegible]	MIDDLE	
Listening	[illegible]	33	[illegible]	[illegible]	[illegible]	HIGH	
Thinking Skills	190	[illegible]	[illegible]	[illegible]	65.6	HIGH	
Basic Battery	322	2[illegible]	[illegible]	73-6	62.9	HIGH	
Complete Battery	402	2[illegible]	[illegible]	73-6	62.9	HIGH	

Otis-Lennon School Ability Test®	Number Possible	Number Correct	SAI	Age PR-S	Age NCE	Scaled Score	Natl Grade PR-S	Natl Grade NCE	National Grade Percentile Bands
Total	72	41	103	57-5	53.7	612	62-6	56.4	
Verbal	36	18	100	50-5	50.0	603	54-5	52.1	
Nonverbal	36	23	106	65-6	58.1	622	70-6	61.0	

FIGURE 2 Display containing many overall scores (display from a commercial test publisher's score report). Reproduced from the *Stanford Achievement Test: Tenth Edition.* Copyright © 2003 by Harcourt Educational Measurement, a Harcourt Assessment Company. Reproduced by permission. All rights reserved.

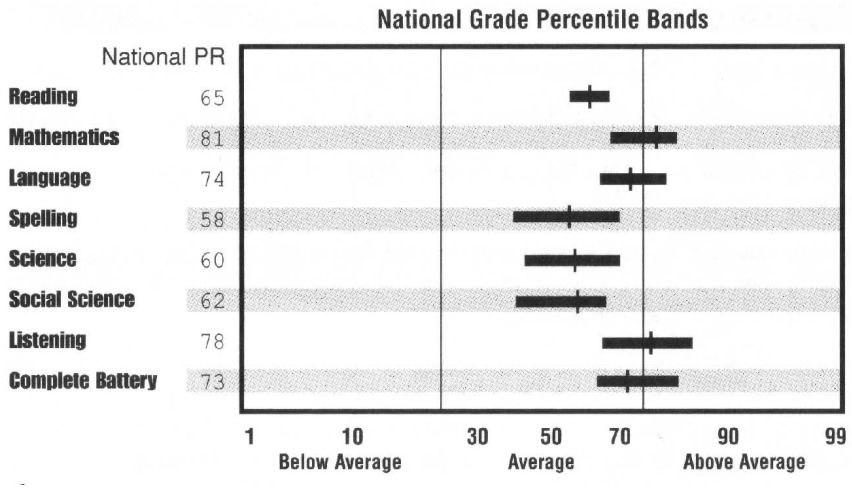


FIGURE 3 Overall scores reported for simplified score report (display from a commercial test publisher's score report). Reproduced from the *Stanford Achievement Test: Tenth Edition.* Copyright © 2003 by Harcourt Educational Measurement, a Harcourt Assessment Company. Reproduced by permission. All rights reserved.

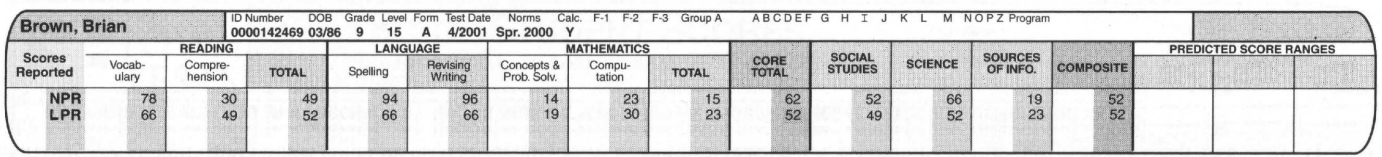
(as shown in Figure 10, one state supplemented this strategy with illustrative test items); and providing a narrative summary and interpretation of the overall scores obtained by a student (see Figure 11).

Overall Results in Relation to Performance Levels

Ten of the 11 states and both Canadian provinces reported students' overall results in relation to state or provincial performance levels (the remaining state reported its results in relation to state content standards, with each individual school district determining what standards were required for graduation). Although none of the commercial test reports reviewed in this study reported student results in relation to performance levels, it is clear from the NEGP (1998) study and a review of commercial test publishers' promotional material that at least two of the test publishers do report results in this manner on other score reports.

All states and Canadian provinces that reported overall results in relation to performance levels displayed them in multiple ways (e.g., numerically, graphically, using text). Figures 7, 12, 13, and 14 illustrate the ways three states and one Canadian province reported overall results in relation to performance levels.

Six states provided general or detailed descriptions of relevant performance levels on their student score reports. Examples of detailed and general descriptions are provided in Figures 14 and 15, respectively.



Brown, Brian	ID Number	DOB	Grade	Level	Form	Test Date	Norms	Calc.	F-1	F-2	F-3	Group A	A B C D E F G H I J K L M N O P Z Program
	0000142469	03/86	9	15	A	4/2001	Spr. 2000	Y					

Scores Reported	READING			LANGUAGE		MATHEMATICS			CORE TOTAL	SOCIAL STUDIES	SCIENCE	SOURCES OF INFO.	COMPOSITE	PREDICTED SCORE RANGES		
	Vocab-ulary	Compre-hension	TOTAL	Spelling	Revising Writing	Concepts & Prob. Solv.	Compu-tation	TOTAL								
NPR	78	30	49	94	96	14	23	15	62	52	66	19	52			
LPR	66	49	52	66	66	19	30	23	52	49	52	23	52			

FIGURE 4 Overall scores presented numerically (display from a commercial test publisher's score report). Reproduced from the *Iowa Tests of Educational Development*® (University of Iowa, 2001b). Copyright © 2001 by The University of Iowa. Used with permission of the publisher. All rights reserved. No part of this work may be reproduced without the prior written permission of The Riverside Publishing Company. Address inquiries to Permissions, The Riverside Publishing Company, 425 Spring Lake Drive, Itasca Illinois 60143.

Name: **District:** **School:**

Highlights: In Mathematics, your scaled score of 1366 indicates achievement at the Proficient Performance Level. You scored as high or higher than 59 percent of 11th grade students. You performed better in Computation and Estimation than in Mathematical Problem Solving.

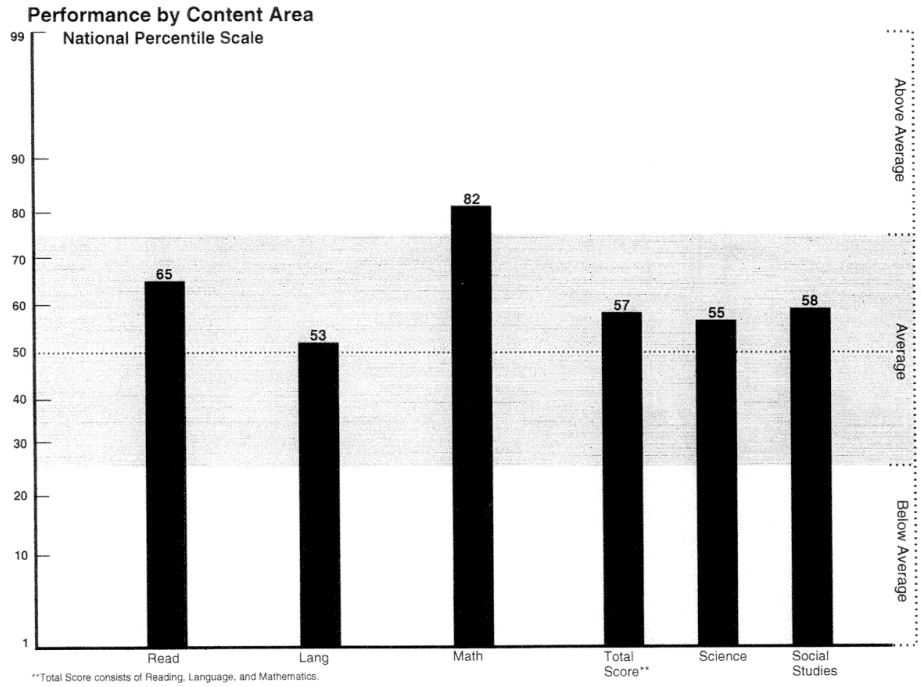
In Reading, your scaled score of 1358 indicates achievement at the Proficient Performance Level. You scored as high or higher than 55 percent of 11th grade students. You performed better in Reading Independently than in Research.

Total Mathematics and Reading Results

Two dashes (- -) are reported if you did not participate in the assessment (see "Highlights").

Content Area	Scaled Score	Standard Error	Performance Level Achieved	Percent of Students Statewide Scoring at Each Performance Level			
				Below Basic	Basic	Proficient	Advanced
Mathematics	1366	50	Proficient	29.0	21.4	26.9	22.7
Reading	1358	61	Proficient	19.7	21.3	43.3	15.7

FIGURE 5 Overall results from a state's score report.



Observations

The height of each bar shows your student's National Percentile score on each test. The percentile scale is shown on the left. The graph shows that your student achieved a National Percentile of 65 in Reading. This means your child scored higher than approximately 65 percent of the students in the nation.

The scale on the right side of the graph shows the score ranges that represent average, above average, and below average in terms of National Percentiles. Average is defined as the middle 50 percent of students nationally, consisting of the 25th through the 75th National Percentiles. Your student has five out of six scores in the average range, shown as a gray horizontal band in the middle of the graph. One score is in the above-average range and no scores are in the below-average range.

See the back of this page for content area descriptions of the kinds of knowledge, skills, and abilities assessed on the achievement test.

FIGURE 6 Overall scores reported numerically and graphically in a commercial test publisher's score report. Reproduced from *TerraNova* (CTB/McGraw-Hill, 2001b). Copyright © 1997 by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc. Reproduced with permission. All rights reserved.

Three states and one Canadian province explicitly reported results for at least one comparative group in relation to performance levels. This was done by reporting the average score of relevant comparison groups in relation to performance levels (see the top portion of Figure 16) or by reporting the percentage of students from various comparison groups who scored within a particular performance level (see the bottom portion of Figure 16).

Three states included information about the precision of overall results that were reported in relation to performance levels. Figure 7 illustrates how one state reported this information.

I. How did do on these tests?

SUBJECT AREA	PERFORMANCE LEVEL	SCALED SCORE	DISPLAY OF SCORE AND PROBABLE RANGE OF SCORES
			Failing / Needs Improvement / Proficient / Advanced
English Language Arts	Advanced	280	
Mathematics	Proficient	246	

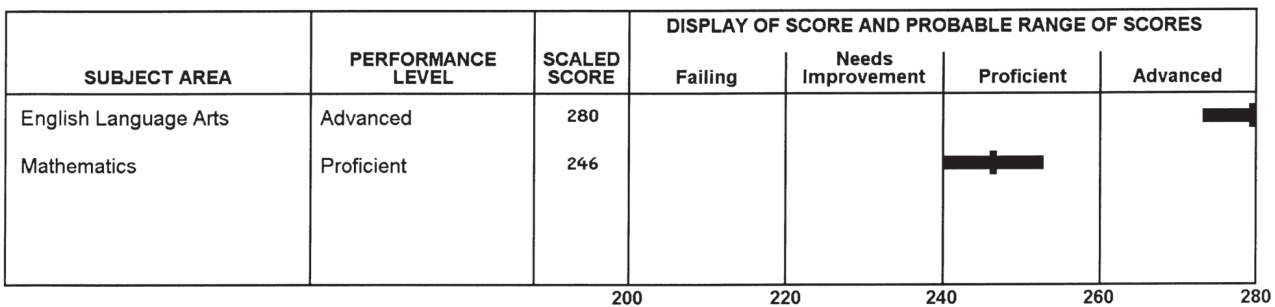


FIGURE 7 Student results with information regarding the precision of overall scaled scores (display from a state's score report).

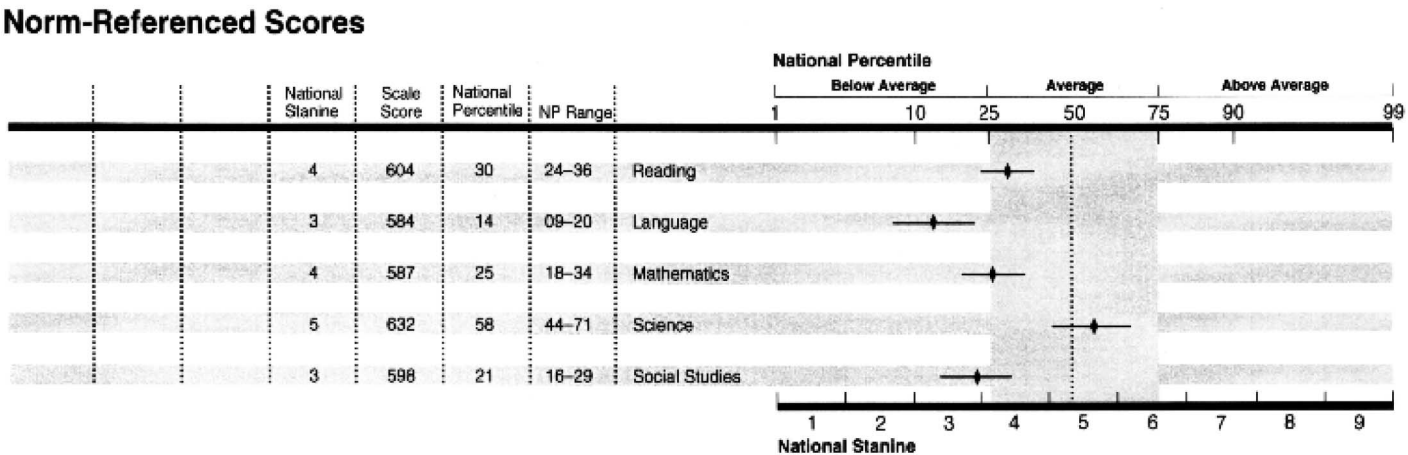


FIGURE 8 Student results with information regarding the precision of overall percentile scores (display from a state's score report). Reproduced from *Wisconsin Student Assessment System* (Wisconsin Department of Public Instruction, 2002). Copyright © 1997 by CTB/ McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc. Reproduced with permission. All rights reserved.

IV. How did your child's score compare to school, district, and state average scores?

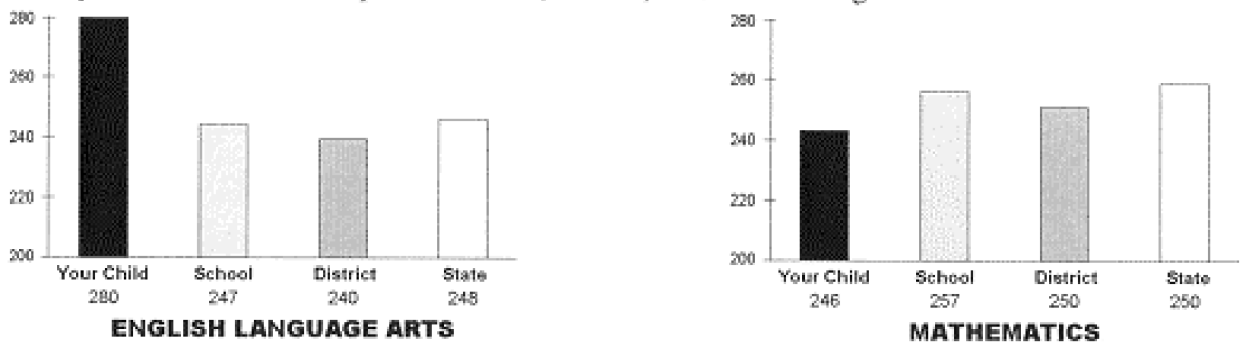


FIGURE 9 Student's overall score in relation to school, district, and state average scores (display from a state's score report).

Sample Test Items

The diagram below gives you an idea of what the total scores mean. The line shows the range of scores. Each box contains a description of some of the skills shown by typical students with scores around the level shown and an illustrative test item.

0 — 17 — 35 — 52 — 70

Students at this level may:

- Apply computational skills when given a procedure
- Use spatial visualization
- Solve problems requiring a single step
- Use basic arithmetic skills

Sample Item:

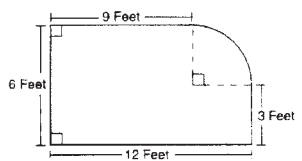
A rabbit population grew in the following pattern: 2, 4, 8, 16,... If all the rabbits live and the pattern continues, how many rabbits will be in the 8th generation?

A. 32
B. 128
C. 256
D. 512

Students at this level may:

- Apply computational skills when given a formula
- Select and correctly apply a formula to solve a problem
- Solve problems requiring two or three steps
- Use basic mathematical concepts

Sample Item:



Find the area of the figure. Use 3.14 for π.

E. 67.71 square feet
F. 70.07 square feet
G. 72 square feet
H. 91.26 square feet

Students at this level may:

- Apply basic concepts and skills in unfamiliar situations
- Create an expression or equation to represent a situation
- Solve problems requiring multiple steps
- Justify reasoning

Sample Item:

A mathematical diversion sometimes used to amaze and amuse people is the creation of a "black hole" for a certain number. You ask your subjects to take any number, go through some manipulations, and then you "magically" tell them the remainder. Here is one such "black hole" for the number 3:

Pick any number
Multiply it by 20
Add 15 to your result
Divide that result by 5
Subtract 4 times the number you picked
The answer is always 3.

Show that this will work with any positive number. Explain your answer using algebraic expressions.

FIGURE 10 Skills and knowledge typically possessed by students with a particular overall score, enhanced by sample test items (display from a state's score report).

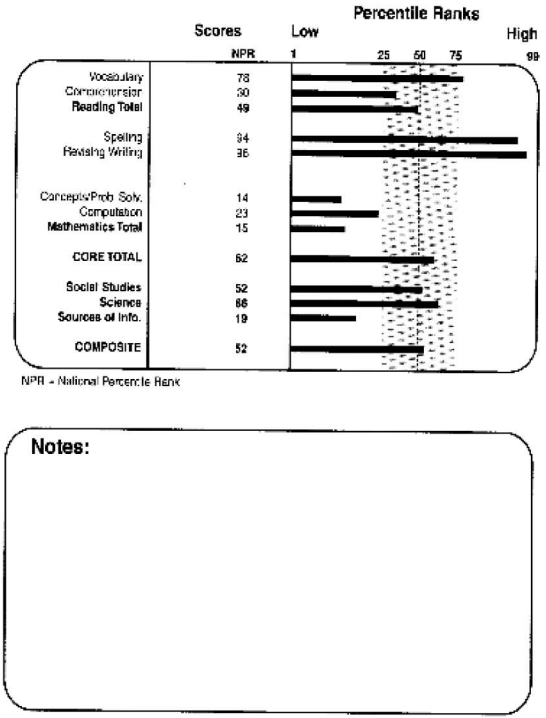


FIGURE 11 Narrative summary and interpretation of the overall scores obtained by a student (display from a commercial test publisher's score report). Reproduced from the *Iowa Tests of Educational Development*® (University of Iowa, 2001a). Copyright © 2001 by The University of Iowa. Used with permission of the publisher. All rights reserved. No part of this work may be reproduced without the prior written permission of The Riverside Publishing Company. Address inquiries to Permissions, The Riverside Publishing Company, 425 Spring Lake Drive, Itasca, Illinois 60143.

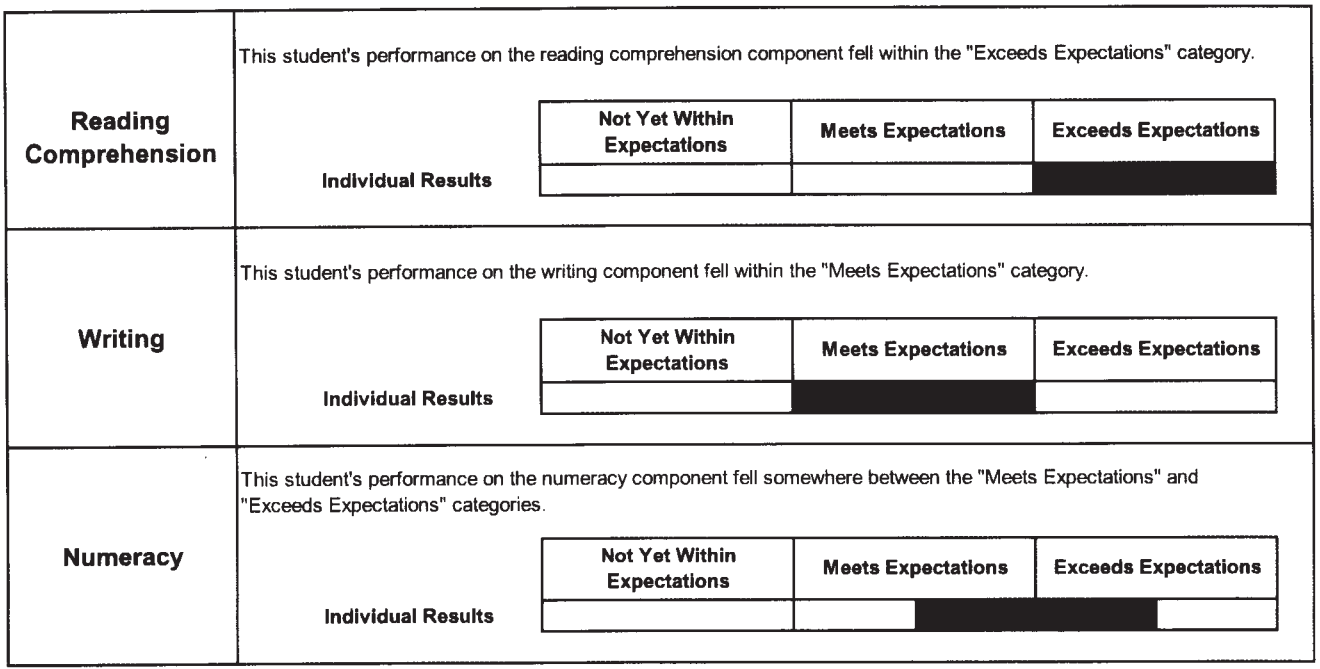


FIGURE 12 Reporting student results in relation to performance levels using text and graphics (display from a provincial score report).

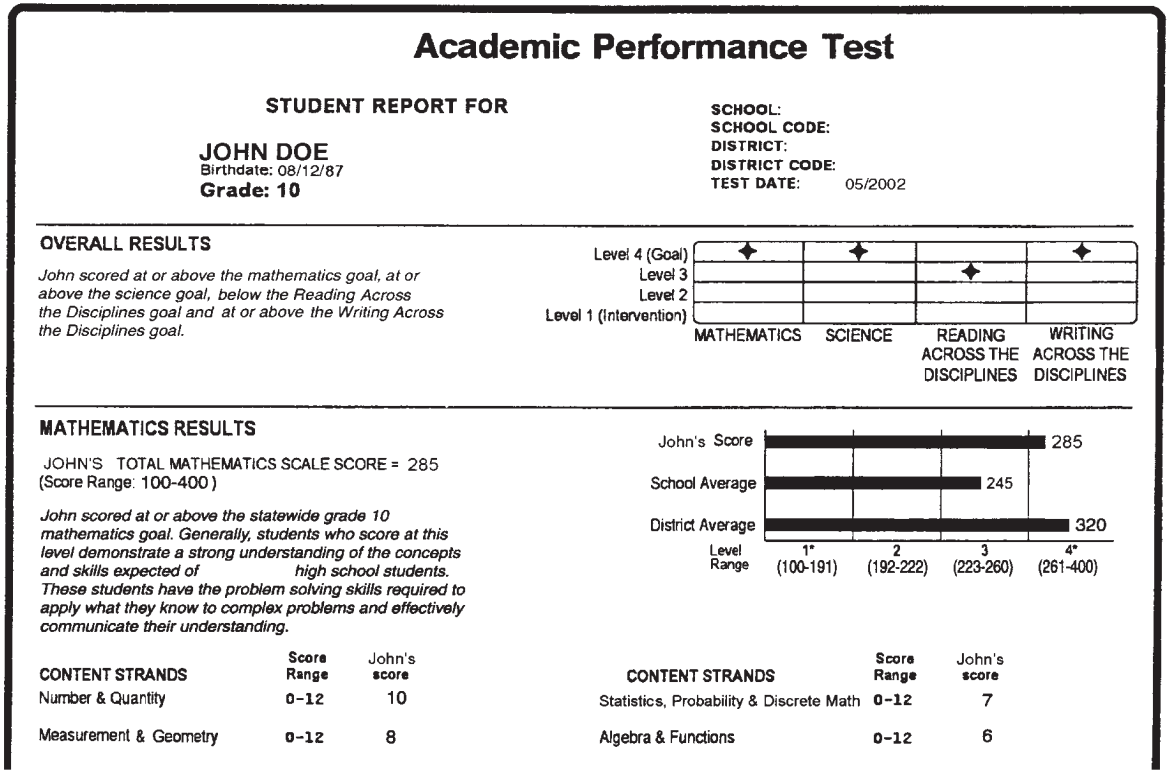


FIGURE 13 Overall results in relation to performance levels, accompanied by comparative and skill-based information (display from a state's score report).

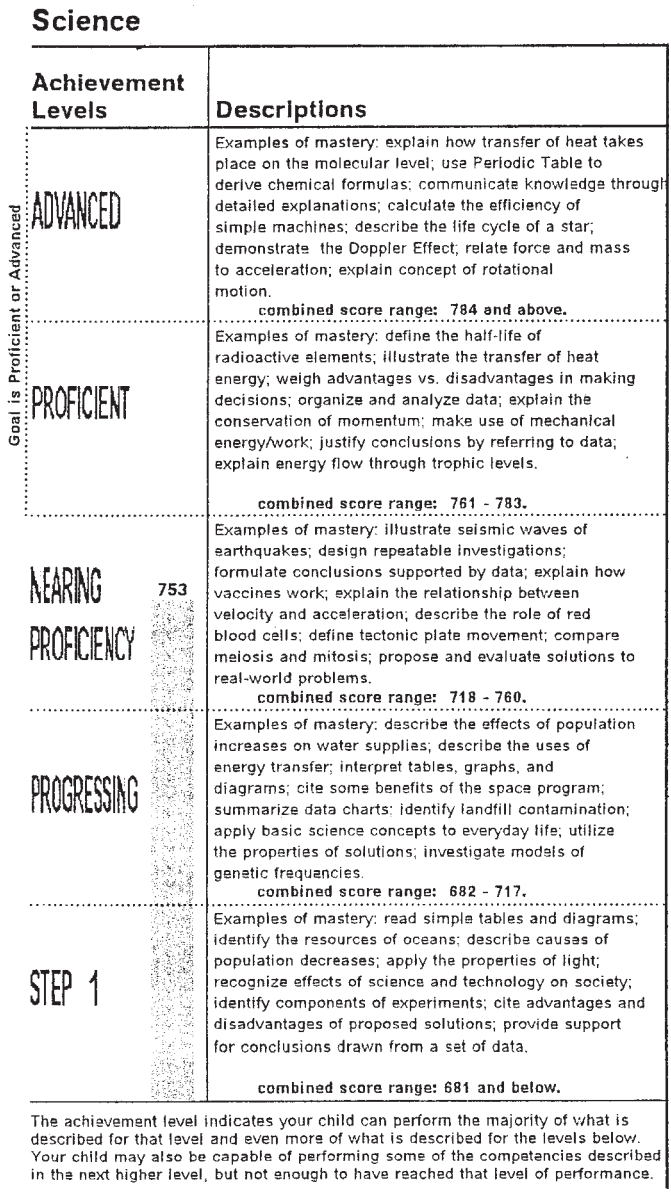


FIGURE 14 Detailed description of performance levels (display from a state's score report). Reproduced from *Missouri Assessment Program* (CTB/McGraw-Hill, 1997). Copyright © 1997 by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc. Reproduced with permission. All rights reserved.

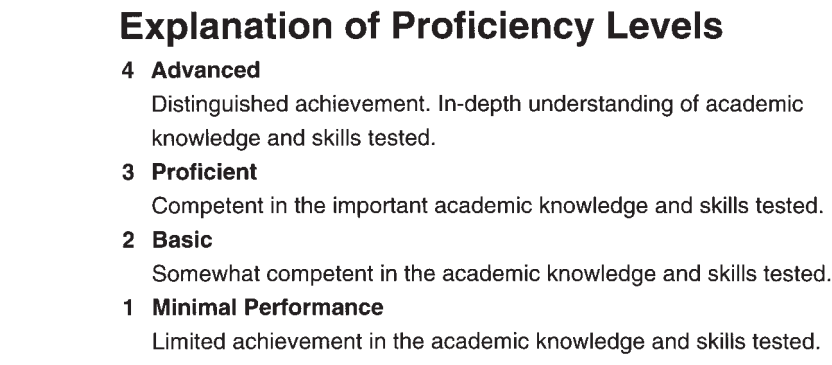


FIGURE 15 General descriptions of performance levels (display from a state's score report). Reproduced from *Wisconsin Student Assessment System* (Wisconsin Department of Public Instruction, 2002). Copyright © 1997 by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc. Reproduced with permission. All rights reserved.

Diagnostic Information

In this study, diagnostic information was operationally defined as information that was more detailed than what was reported at the general subject area level (e.g., general descriptions of various skills that make up subject-level performance were not considered to be diagnostic for our purposes). Across reports, two general types of diagnostic information were identified: (a) student results by subdomain categories (i.e., skills or content areas within a particular subject area) and (b) specific skills or knowledge a student demonstrated on the assessment or should develop to improve his or her performance.

All 11 states and all 3 commercial test publishers included some type of diagnostic information on their student reports (the home reports for 2 of the publishers did not include this information, however). One of the 2 Canadian provinces reported diagnostic information, but only for those students who did not pass the relevant portion of the test. The different ways diagnostic information was reported are summarized in the following sections.

Subdomain results. Eight of the 11 states, 1 of the 2 Canadian provinces, and all 3 commercial test publishers reported subdomain results in their student score reports. These results were typically reported as raw scores, percent correct scores, or percentile rank scores. A student's relative strengths and weaknesses within a particular subject area could be identified by comparing and interpreting these subdomain results.

Subdomain results were typically reported in a numerical manner (one state, however, reported subdomain results only in a graphical manner). Three states and three commercial test publishers reported subdomain results in multiple ways

SECTION 1: STUDENT/DISTRICT/STATE TEST PERFORMANCE

Results from the Graduation Exit Examination for the test, are reported in terms of achievement levels. As shown below, your English Language Arts achievement level is **Unsatisfactory.** Your score of **245** indicates that you **did not pass** this test. Your performance is **lower than** the district average and **lower than** the state average.

	Achievement Level				
	UNSATISFACTORY	A.B.	BASIC	PROF.	ADVANCED
	100–269	270–298	299–346	347–397	398–500
AARON'S SCORE (245)	↑				
District Average (308)			↑		
State Average (298)			↑		

For your school district's students, 2% performed at the Advanced level, 16% at the Proficient level, 42% at the Basic level, 22% at the Approaching Basic level, and 18% at the Unsatisfactory level. State results are also shown below.

Achievement Level	Description	District Percent	State Percent
	A student at this level:		
Advanced	has demonstrated superior performance beyond the proficient level of mastery	2%	1%
Proficient	has demonstrated competency over challenging subject matter; well-prepared for the next level of schooling	16%	13%
Basic	has demonstrated only the fundamental knowledge and skills needed for the next level of schooling	42%	38%
Approaching Basic	has only partially demonstrated fundamental knowledge and skills needed for the next level of schooling	22%	23%
Unsatisfactory	has not demonstrated the fundamental knowledge and skills needed for the next level of schooling	18%	24%

The percent of students in the district and state across achievement levels may not add to 100% due to rounding.

FIGURE 16 Results of comparison groups in relation to performance levels (display from a state's score report).

(e.g., numerically and graphically). Two examples of how subdomain results were reported are shown in Figures 17 and 18.

No state or Canadian province provided information about the precision of reported subdomain scores. Reports from two of the commercial test publishers included confidence intervals for subdomain scores. Figures 2 and 19 illustrate ways the precision of subdomain scores was depicted in these commercial test reports.

No state or Canadian province reported subdomain scores in relation to expected levels of student performance, although one state provided a general benchmark for evaluating subdomain performance by reporting average subdomain scores for "proficient" students. A report from one commercial test publisher compared students' subdomain scores with a range of scores representing moderate mastery (see Figure 19).

Four of the 11 states and all 3 commercial test publishers reported subdomain scores in relation to the performance of other students. Neither of the Canadian provinces reported this type of information.

In three states, student performance on each subdomain was compared with the average performance of all students in the state (see Figure 20), with the average performance of students in the state with the same reported history of instruction (see Figure 21), or, as described earlier, with the average performance of students who obtained a "proficient" score on the overall test. In one state, student subdomain scores were reported graphically as a state percentile rank. The three commercial test publishers reported comparative information about student performance by subdomains in terms of national (and in one report, local) percentiles, stanines, or average percent correct scores.

Specific skills or knowledge demonstrated or to be developed. A second approach for reporting diagnostic information was to list a particular student's specific strengths or weaknesses within a given subject area. This information, which was more specific than the generic descriptions used to provide meaning to subdomain or overall results, was included in reports from two states and one Canadian province. Figures 22 and 23 show how the two states reported particular strengths or weaknesses of individual students.

Types of Information Included in the Interpretive Guides and Methods of Reporting

Answers to General Questions About the Assessment

Interpretive material that accompanied or was an integral part of student score reports usually answered one or more key questions parents and guardians might have about the assessment or the assessment results (e.g., What was the purpose of the test? What was assessed? What did the test look like? Where can parents get

Feedback on Reading Skills

Students were asked to complete a total of 12 reading selections, divided into three different types:

- Information (e.g., explanation, opinion)
- Graphic (e.g., graph, schedule, instructions)
- Narrative (e.g., story, dialogue)

The reading skills assessed were

- Understanding directly stated ideas and information
- Understanding indirectly stated ideas and information
- Making connections between personal experiences and the ideas and information in the reading selections

Reading Scores

Reading skills were scored 0 points for an incorrect answer, 1 point for a partially correct answer (where possible) and 2 points for a correct answer.

Reading Skills	Reading Selection Types		
	Information	Graphic	Narrative
Understands directly stated ideas and information	/ 30	/ 16	/ 14
Understands indirectly stated ideas and information	/ 40	/ 22	/ 28
Makes connections between personal experiences and the ideas and information in the reading selections	/ 20	/ 12	/ 18

Your total reading score is points.

Your score was calculated by adding the total number of points for questions marked as correct (2 points each) to the total for those marked partially correct (1 point each). The score to pass reading, based on the provincial standard-setting process for the , was points or higher. (This score varies if you wrote a special version of the .)

FIGURE 17 Subdomain results from a provincial score report.

SECTION 2: INDIVIDUAL STUDENT PERFORMANCE BY CONTENT STANDARD

The English Language Arts test measures concepts and skills in six of seven areas that are referred to as content standards. The seven content standards specify what students are expected to know and be able to do in English language arts. The graph below shows how many points you received for each content standard.

Content Standard	Score Points	Percent Correct
1. Read, comprehend, and respond	1 of 10	10%
2. Write competently	6.0 of 8	75%
3. Use conventions of language	9.0 of 12	75%
4. Apply speaking/listening skills	not assessed	
5. Locate, select, and synthesize information	4 of 9	44%
6. Read, analyze, and respond to literature	3 of 8	38%
7. Apply reasoning and problem solving skills	4.0 of 22	18%

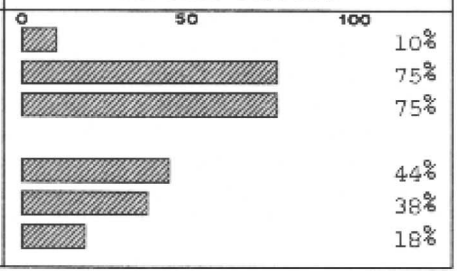


FIGURE 18 Subdomain results from a state's score report.

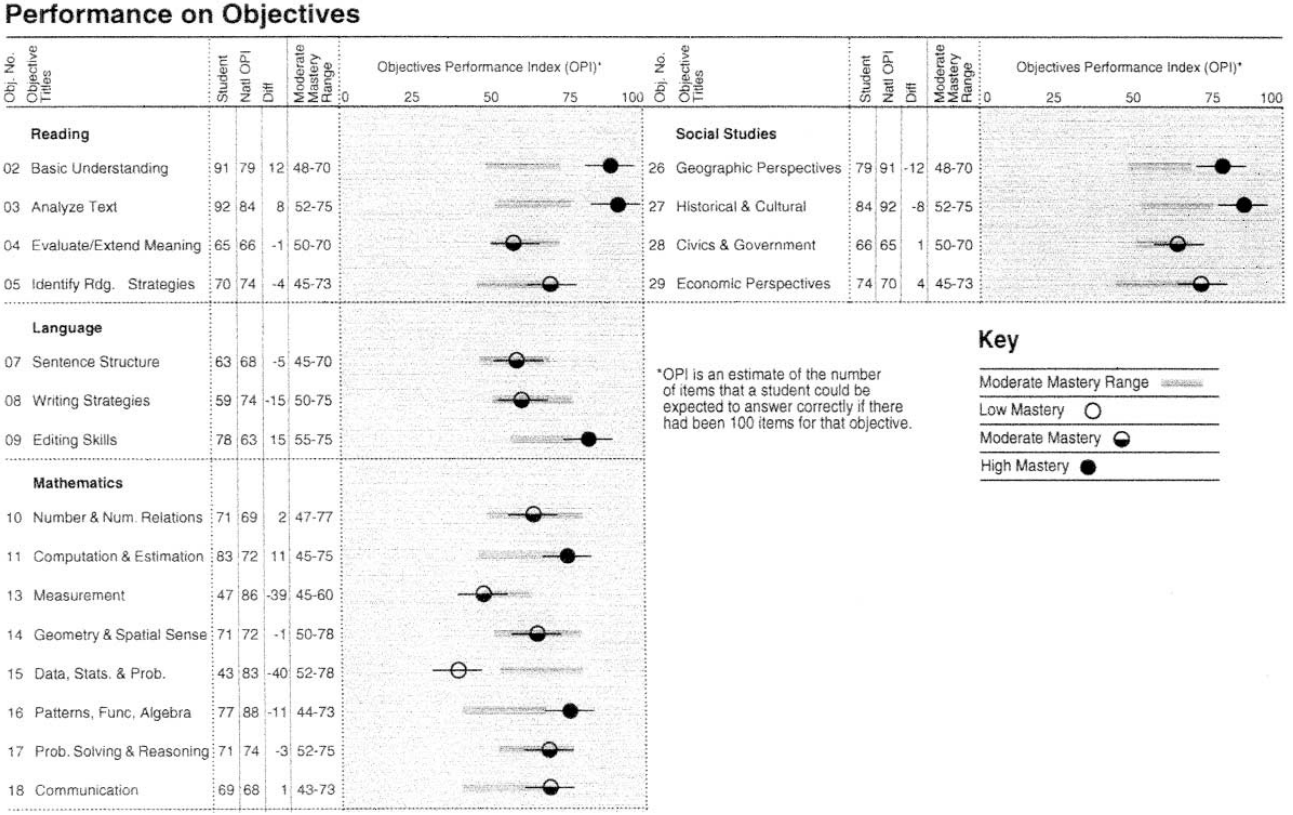


FIGURE 19 Information regarding the precision of subdomain scores (display from a commercial test publisher's score report). Reproduced from *TerraNova* (CTB/McGraw-Hill, 2001c). Copyright © 1997 by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc. Reproduced with permission. All rights reserved.

Results by Academic Standards

To assist in understanding how well you performed in each standard, the following results are presented:

- Points achieved, maximum points possible and average points achieved for students statewide
- A graph that shows the difference between your points achieved and the state average
- Two dashes (- -) are reported if you did not participate in the assessment (see "Highlights").

Because of differences in instructional focus that may occur from grade to grade, the number of questions and possible points differ across Academic Standards. The number of questions measuring an Academic Standard is small compared with the total number of questions. For this reason, the Academic Standards with the most possible points are usually measured more accurately.

Results for Mathematics Academic Standards

Mathematics Academic Standards	Points Achieved	Points Possible	State Average	Difference Between Points Achieved and State Average (-10 -8 -6 -4 -2 0 2 4 6 8 10)
2.1 Number Systems and Relationships	1	5	2.4	
2.2 Computation and Estimation	10	10	6.7	
Without a Calculator	5	5	3.8	
With a Calculator	5	5	2.9	
2.3 Measurement and Estimation	4	10	5.2	
2.4 Mathematical Reasoning	4	5	3.3	
2.6 Statistics and Data Analysis	9	10	6.9	
2.7 Probability and Predictions	2	5	2.3	
2.8 Algebra and Functions	9	19	9.6	
2.9 Geometry	5	11	4.8	
2.10 Trigonometry	5	5	2.8	
2.11 Concepts of Calculus	2	5	2.9	
2.5 Mathematical Problem Solving *	2	15	4.6	
TOTAL MATHEMATICS	51	85	48.9	

* All items for Academic Standard 2.5, *Mathematical Problem Solving*, are also included in other Academic Standards. *Total Mathematics* does not include points listed under Academic Standard 2.5.

Results for Reading Academic Standards

Reading Academic Standards	Points Achieved	Points Possible	State Average	Difference Between Points Achieved and State Average (-10 -8 -6 -4 -2 0 2 4 6 8 10)
1.1 Reading Independently	16	16	12.2	
1.2 Reading Critically	11	16	10.3	
1.3 Analyzing/Interpreting Literature	15	26	16.0	
1.7 English Language Characteristics	9	12	7.4	
1.8 Research	10	16	11.1	
TOTAL READING	51	88	57.0	

Percentages and state averages are rounded. Therefore, "Total Mathematics" and "Total Reading" may vary slightly from the sum of the parts.

FIGURE 20 Subdomain results in relation to average points achieved by student in the state (display from a state's score report).

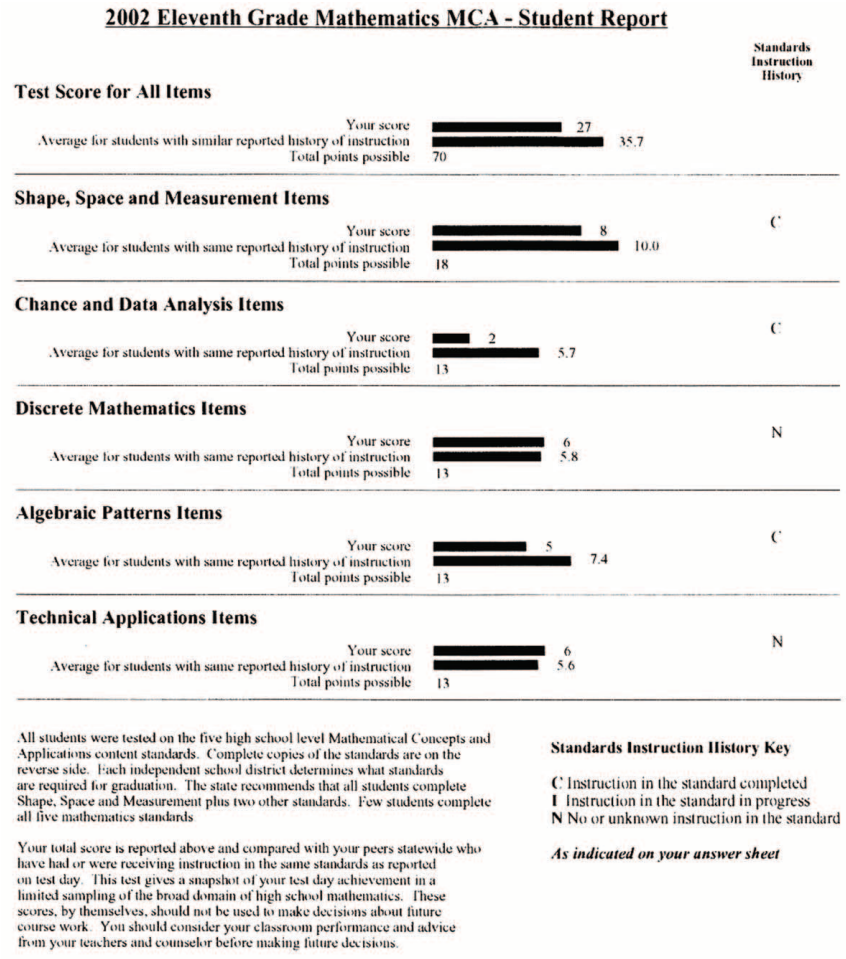


FIGURE 21 Subdomain results in comparison to students in the state with the same reported history of instruction (display from a state's score report).

more information? What can parents and guardians do to help students improve?). Several states reported this information in the form of a letter to parents from a state official (see Figure 24). One state included a table of contents in its interpretive guide to facilitate retrieval of this information.

Purpose of assessment. Information regarding the purpose of the assessment was included in the interpretive material of nine states (one of these states provided this information only in its Web-based interpretive guide), both Canadian

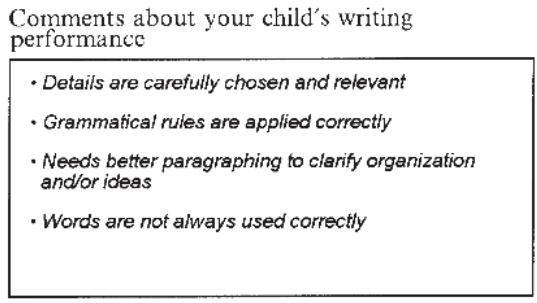


FIGURE 22 Particular strengths and weaknesses of an individual student (display from a state's score report).

provinces, and two commercial test publishers. An example of how one state described the purpose of its assessment is provided in Figure 25.

Content assessed. All states, Canadian provinces, and commercial test publishers provided some information about what was assessed by the test (i.e., beyond general subject areas). The level of detail ranged from general descriptions of subdomain reporting categories (e.g., number and number sense, measurement and geometry) to complete descriptions of each relevant content standard (see Figure 26).

What the test looked like. Eight of the 11 states, both Canadian provinces, and 2 of the 3 commercial test publishers provided information about what the test looked like (1 of the 7 states provided information about what the test looked like on its Web-based interpretive guide but not on its paper-based guide). Six states, both Canadian provinces, and 2 commercial test publishers provided general descriptions of the types of questions that made up the test (e.g., multiple choice and constructed response). Figure 27 illustrates how 1 state described the types of items used in its assessment. Only 2 states and 1 commercial test publisher included sample questions in their interpretive guides (1 of these states included sample questions only in its Web-based interpretive guide). Figure 28 illustrates how 1 commercial test publisher displayed sample questions in its interpretive guide. Two states identified separate resources that contained actual test items.

Suggestions to improve performance. Five states, one province, and all three commercial test publishers provided explicit suggestions about what parents, guardians, or the students themselves could do to improve student performance. These suggestions ranged from engaging in general activities that can

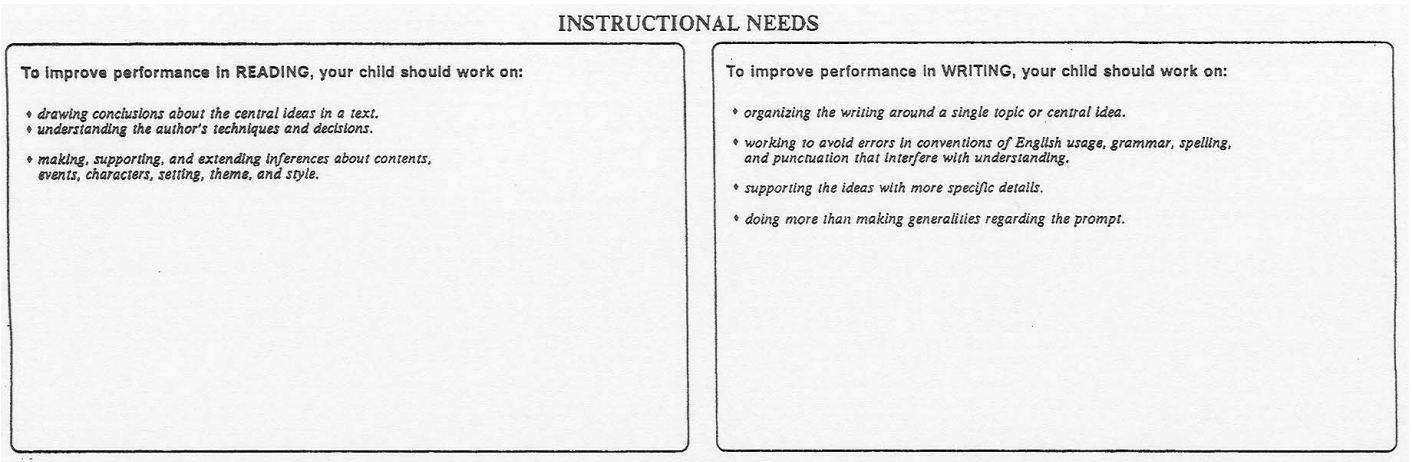


FIGURE 23 Specific instructional needs for an individual student (display from a state's score report).

Dear Parents/Guardians:

The assesses students' progress relative to the Standards. The is the required assessment program for all public-school children in the subject areas of mathematics (grades 4, 8, and 10), science (grades 3, 7, and 10), communication arts (grades 3, 7, and 11), and social studies (grades 4, 8, and 11).

Most assessments include three types of items:

- The **multiple-choice** component is composed of custom selected-response items and/or the survey portion of a nationally normed test.
- **Constructed-response** items require students to supply (rather than select) an appropriate response.
- **Performance events** are longer, more demanding tasks requiring students to work through problems, experiments, arguments, or extended pieces of writing.

Achievement-level scores provide a measure of what students can do in terms of the content and skills assessed by the , which are typically found in the curriculum for the grade span being assessed. "Proficient" or "Advanced" levels of achievement are desirable.

The height of the vertical bar indicates the level your child reached at the time of testing. The bar is always positioned in the middle of the level, but your child's actual score is printed at the TOP of the bar. The score range possible for each level is printed in the "Descriptions" text, so you may compare your child's score to the range for the level achieved. Your child may have just reached the level indicated by the bar or be very close to moving on to the next-higher level. The achievement level attained by your child indicates he/she can perform the majority of what is described for that level and even more of what is described for the levels below. Your child may also be capable of performing some of the competencies described in the next-higher level, but not enough to have reached that level of performance.

Look at the skills and knowledge described in the next-higher level. These are the competencies your child needs to demonstrate to show academic growth. If your child is at the "Advanced" level, check with your child's teacher for enrichment activities.

The **national percentile** represents the percentage of students in the norm group whose scores fall below a given student's score. For example, a student whose NP is 65 scored higher than 65 percent of the students in the norm group.

Standards are assessed statewide and in the classroom. There are 40 content standards that provide a solid foundation of knowledge and basic skills every student should acquire in mathematics, science, communication arts, social studies, health/physical education, and fine arts. These standards define the body of knowledge that every child should experience within the K-12 curriculum. There are also 33 process standards that include skills students should master in order to successfully gather, analyze, and apply information; communicate effectively; recognize and solve problems; and become responsible citizens. Not all process standards are assessed and reported.

While intended to establish higher expectations for all of public-school students, the 73 Standards do not represent everything a student should or will learn. However, graduates who met these standards should be well prepared for further education, work, and civic responsibilities.

We hope this report helps you gain insight into your child's academic achievement.

FIGURE 24 Letter to parents/guardians included on the reverse side of a state's score report. Reproduced from *Missouri Assessment Program* (CTB/McGraw-Hill, 1997). Copyright © 1997 by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc. Reproduced with permission. All rights reserved.

help improve performance in a given subject area (e.g., use home projects to investigate relationships between units used to measure length, area, and volume), to developing the skill set representative of a higher performance level than the one attained by the student (e.g., students at the "basic" level should work on developing identified skills that reflect proficient performance), to working on specific skills identified on the student report that were tailored to the performance of individual students (e.g., skills such as those reported in Figures 22 and 23).

Where to get more information. Seven states, both Canadian provinces, and two of the three commercial test publishers provided recommendations about

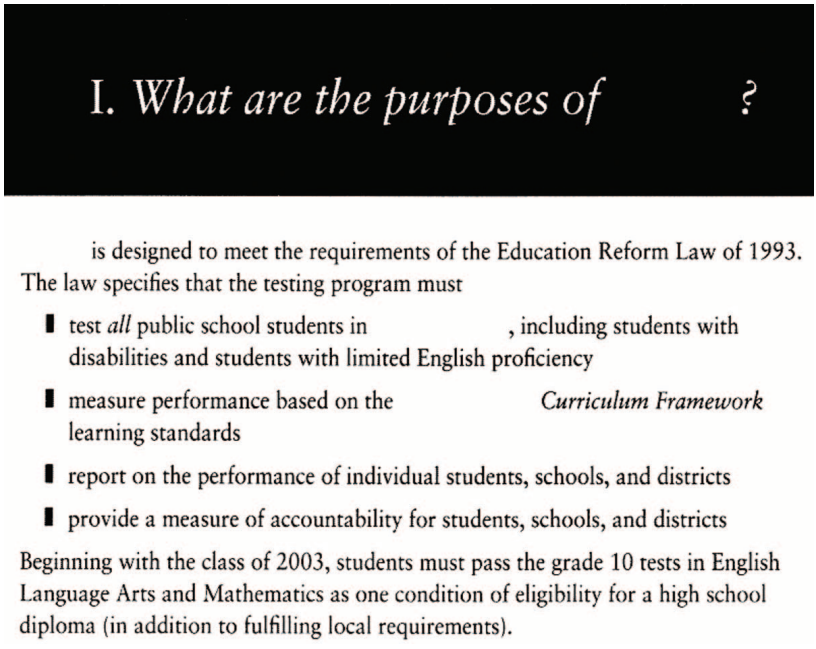


FIGURE 25 Purposes of a state assessment provided in an interpretive guide.

where parents should go to get more information about the test or students' results. In these cases, parents were advised to talk to their child's teacher, guidance counselor, or principal or to review publications and Web sites listed in the interpretive guide.

Details Regarding Key Elements of the Student Reports

All of the interpretive material reviewed provided at least some details about key elements of the student reports. These included descriptions of the relevant sections of the reports as well as definitions for technical terms.

Descriptions of the sections of the reports. All states, Canadian provinces, and commercial test publishers produced interpretive material that described various sections of the student reports. The sections of the reports were described in two general ways, using either descriptive text only or a combination of descriptive text and graphic displays of sample reports. Seven states, both Canadian provinces, and two commercial test publishers provided only written descriptions of relevant sections (see Figure 29). Four states and one commercial test publisher described various sections of the reports using a combination of descriptive

High Standards in Mathematical Concepts and Applications

The state content standards are clearly defined expectations against which individual student achievement and progress may be judged. They outline what a student needs to know and do in a particular subject. Local public school districts determine how the content standards are taught and how student achievement is assessed. The 11th Grade Mathematics MCA gives a statewide glimpse of student achievement in the following five standards.

Shape, Space, and Measurement

A student shall:

A. demonstrate understanding of the characteristics of geometric figures in both two and three dimensions, including reflections, rotations, and translations; congruence and similarity; perimeter, area, and volume; distance; scaling; and symmetry;
B. use spatial visualization to model geometric structures and solve problems;
C. analyze characteristics of shape, size, and space in art, architecture, design, or nature;
D. translate between numerical relationships and geometric representations to analyze problem situations, scale models, or measurement;
E. use properties of shape, location, or measurement to justify reasoning in a logical argument; and
F. demonstrate understanding of measurement accuracy, error, and tolerances.

Chance and Data Analysis

A student shall:

A. demonstrate understanding of the statistical concepts of measures of center, variability, and rank; differences between correlation and causation; sampling procedures; line or curve of best fit; and concepts related to uncertainty of randomness, permutations, combinations, and theoretical and experimental probabilities;
B. investigate a problem of significance by formulating a complex question, designing a statistical study, collecting data, representing data appropriately, using appropriate statistics to summarize data, determining whether additional data and analysis are necessary, drawing conclusions based on data, and communicating the results appropriately for the intended audience;
C. analyze and evaluate the statistical design, survey procedures, and reasonableness of conclusions in a published study or article;
D. use probability experiments, simulations, or theory-to-model situations involving uncertainty; and
E. make predictions based on the model.

Discrete Mathematics

A student shall use discrete structures to demonstrate mathematical relationships and solve problems by:

A. describing the difference between discrete and continuous models of data and permutations, combinations, and other principles of systematic counting;
B. translating between real-world situations and discrete mathematical models using vertex-edge graphs, matrices, verbal descriptions and sequences;
C. analyzing and modeling iterative and recursive patterns;
D. analyzing and solving problems by building discrete mathematical models, developing and comparing algorithms or sequences of procedures, and determining whether solutions exist, the number of possible solutions, and the best solutions; and
E. using properties of mathematics to justify reasoning in a logical argument.

Algebraic Patterns

A student shall demonstrate the ability to identify rates of change in different models of linear relationships and know characteristics of polynomial, exponential, and periodic functions and relations; functional notation; and terminology by:

A. translating between real-world situations and mathematical models using graphs; matrices; data tables, spread sheets, or both; verbal descriptions; and algebraic expressions;
B. generalizing patterns and building mathematical models to describe and predict real situations including linear, exponential growth and decay, and periodic;
C. using algebraic concepts and processes to represent and solve problems involving variable quantities; and
D. using properties of algebra to justify reasoning using a logical argument.

Technical Applications

A student shall:

A. demonstrate knowledge of computational technologies; how to use complex measurement equipment for several systems; how to convert between measuring systems; how to measure to scale; how to calculate quantities using algebraic formulas; how to read and interpret information in complex graphs, tables, and charts; scientific and exponential notation used in complex systems; trigonometric applications appropriate to technical situations; and fundamental geometric constructions or calculations used in drafting or construction;
B. create a set of plans to design or modify a complex structure, product, or system by researching background information, calculating mathematical-specifications, and developing a materials list that matches mathematical-specifications;
C. construct a complex structure, product, or model to mathematical-specifications; and
D. analyze existing complex structure, product, or system for purposes of maintenance, repair, trouble shooting, or optimizing function.

FIGURE 26 Complete descriptions of content standards (display from the reverse side of a state's score report).

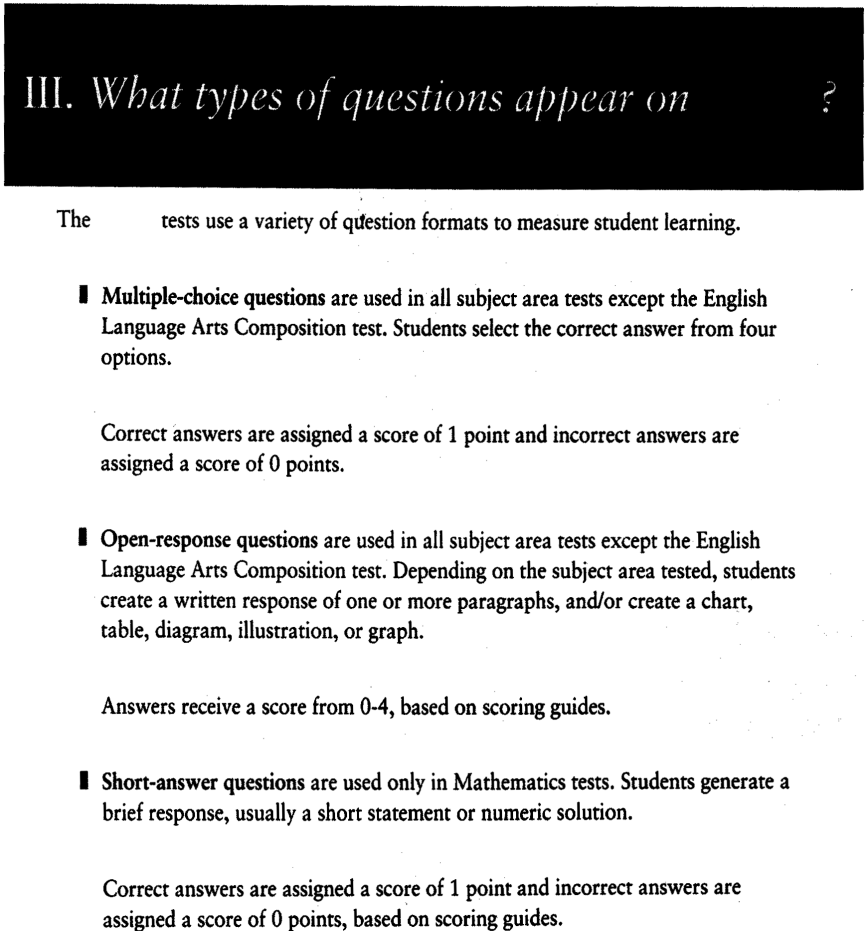
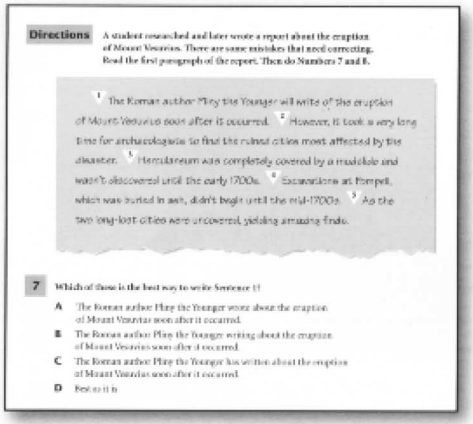


FIGURE 27 Description of the different types of questions used on a state assessment.

text and graphic displays (see Figures 30–32 for examples of how three states used graphics to describe sections of their student reports). One state and one commercial test publisher produced interactive Web-based guides that allow users to get detailed information about different sections of the reports by clicking on the sections of interest.

Definitions for technical terms. Most key terms used in the student reports were defined in the interpretive material supplied by all of the states, Canadian provinces, and commercial test publishers that participated in the study. Three



Directions A student researched and later wrote a report about the eruption of Mount Vesuvius. There are some mistakes that need correcting. Read the first paragraph of the report. Then do Numbers 7 and 8.

[1] The Roman author Pliny the Younger will write of the eruption of Mount Vesuvius soon after it occurred. [2] However, it took a very long time for archaeologists to find the ruined cities most affected by this disaster. [3] Herculaneum was completely covered by a mudslide and wasn't discovered until the early 1700s. [4] Excavations at Pompeii, which was buried in ash, didn't begin until the mid-1700s. [5] As the two long-lost cities were uncovered, yielding amazing finds.

7 Which of these is the best way to write Sentence 1?

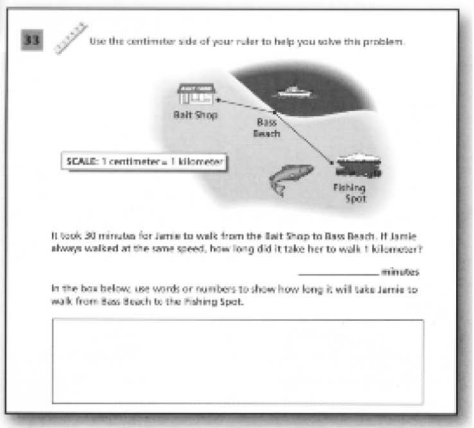
A The Roman author Pliny the Younger wrote about the eruption of Mount Vesuvius soon after it occurred.

B The Roman author Pliny the Younger writing about the eruption of Mount Vesuvius soon after it occurred.

C The Roman author Pliny the Younger has written about the eruption of Mount Vesuvius soon after it occurred.

D Best as it is

Figure 1
This is an example of a Grade 6 Reading/Language Arts multiple-choice question.



33 Use the centimeter side of your ruler to help you solve this problem.

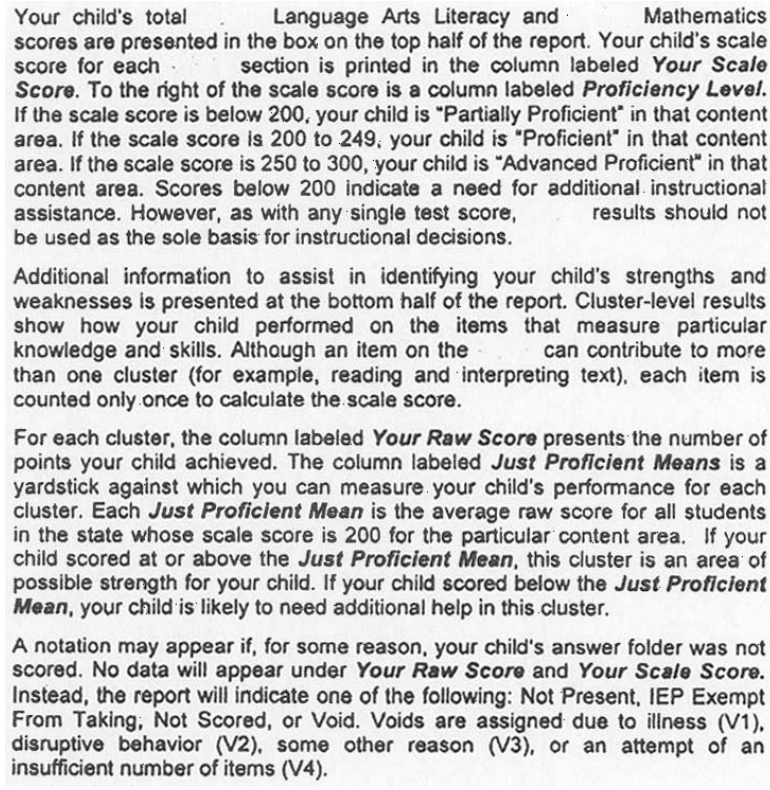
It took 30 minutes for Jamie to walk from the Bait Shop to Bass Beach. If Jamie always walked at the same speed, how long did it take her to walk 1 kilometer?

_____ minutes

In the box below, use words or numbers to show how long it will take Jamie to walk from Bass Beach to the Fishing Spot.

Figure 2
This is an example of a Grade 5 Mathematics constructed-response question.

FIGURE 28 Sample questions included in an interpretive guide from a commercial test publisher. Reproduced from *TerraNova* (CTB/McGraw-Hill, 2001b). Copyright © 1997 by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc. Reproduced with permission. All rights reserved.



Your child's total Language Arts Literacy and Mathematics scores are presented in the box on the top half of the report. Your child's scale score for each section is printed in the column labeled ***Your Scale Score***. To the right of the scale score is a column labeled ***Proficiency Level***. If the scale score is below 200, your child is "Partially Proficient" in that content area. If the scale score is 200 to 249, your child is "Proficient" in that content area. If the scale score is 250 to 300, your child is "Advanced Proficient" in that content area. Scores below 200 indicate a need for additional instructional assistance. However, as with any single test score, results should not be used as the sole basis for instructional decisions.

Additional information to assist in identifying your child's strengths and weaknesses is presented at the bottom half of the report. Cluster-level results show how your child performed on the items that measure particular knowledge and skills. Although an item on the can contribute to more than one cluster (for example, reading and interpreting text), each item is counted only once to calculate the scale score.

For each cluster, the column labeled ***Your Raw Score*** presents the number of points your child achieved. The column labeled ***Just Proficient Means*** is a yardstick against which you can measure your child's performance for each cluster. Each ***Just Proficient Mean*** is the average raw score for all students in the state whose scale score is 200 for the particular content area. If your child scored at or above the ***Just Proficient Mean***, this cluster is an area of possible strength for your child. If your child scored below the ***Just Proficient Mean***, your child is likely to need additional help in this cluster.

A notation may appear if, for some reason, your child's answer folder was not scored. No data will appear under ***Your Raw Score*** and ***Your Scale Score***. Instead, the report will indicate one of the following: Not Present, IEP Exempt From Taking, Not Scored, or Void. Voids are assigned due to illness (V1), disruptive behavior (V2), some other reason (V3), or an attempt of an insufficient number of items (V4).

FIGURE 29 Written descriptions of sections of student score report (taken from the interpretive guide included on the reverse side of a state's score report).

strategies used to communicate this information included descriptive footnotes, definitions embedded in narrative text about the report, and special sections that contained definitions of key terms (e.g., a glossary of key terms was included in the student report of one Canadian province and in the Web-based interpretive guide of one state).

Although commercial test publishers defined all key terms used in their student reports, four states and one Canadian province did not. Three states and one Canadian province did not define the performance levels used to report student-level results. The fourth state did not define an abbreviation included on the score report. One other state did not define all key terms (e.g., standard error) on its paper-based interpretive guide but provided detailed definitions and illustrative examples on its Web-based guide.

Interpreting the Individual Student Report
English Language Arts

REACHING FOR RESULTS is new plan to improve student achievement. A major part of REACHING FOR RESULTS is and the Graduation Exit Examination , the state's new criterion-relerenced tests. Parents, students, and teachers can use the following guide to interpret the information presented on the front of this report.

SECTION 1: STUDENT/DISTRICT/STATE TEST PERFORMANCE

The first table in Section 1 shows a student's performance in terms of achievement level and test score. results are reported according to five achievement levels: Advanced, Proficient, Basic, Approaching Basic, and Unsatisfactory. The student's score, the district's average score, and the state's average score are listed in parentheses on the left side of the table and are also indicated with arrows. A student's performance can be compared to the average performance of students at the district and state levels.

The second table in Section 1 presents the definitions of the five achievement levels and the percentage of students in the district and state at each level. As shown in the sample table below, 1 percent of students in the district performed at the Advanced level, while 2 percent of students in the state were at that level.

(Sample Table)

Achievement Level	Description	District Percent	State Percent
	A student at this level:		
Advanced	has demonstrated superior performance beyond the proficient level of mastery	1%	2%
Proficient	has demonstrated competency over challenging subject matter and is well-prepared for the next level of schooling	9%	12%

SECTION 2: INDIVIDUAL STUDENT PERFORMANCE BY CONTENT STANDARD

The graph in Section 2 reports a student's performance in six of seven areas of English language arts, which are referred to as content standards. The **Score Points** column lists how many points a student received for each content standard. As shown in the sample graph below, "8 of 10" in Standard 1 means this student received 8 points out of 10 possible points available for that standard. The score points may include a decimal because some items are scored by two readers, and an average of the two scores is used. The **Percent Correct** bar graph shows the percentage of points this student received for each standard.

(Sample Graph) (The possible score points shown below are different for each grade level.)

Content Standard	Score Points	Percent Correct
		0 50 100
1. Read, comprehend, and respond	8 of 10	80%
2. Write competently	3.5 of 8	44%

The English Language Arts Content Standards specify concepts and skills that students are expected to know and be able to do. The seven content standards are:

1. Students read, comprehend, and respond to a range of materials using a variety of strategies for different purposes.
2. Students write competently for a variety of purposes and audiences.
3. Students communicate using standard English grammar, usage, sentence structure, punctuation, capitalization, spelling, and handwriting.
4. Students demonstrate competence in speaking and listening as tools for learning and communicating. (not assessed)
5. Students locate, select, and synthesize information from a variety of texts, media, references, and technological sources to acquire and communicate knowledge.
6. Students read, analyze, and respond to literature as a record of life experiences.
7. Students apply reasoning and problem-solving skills to reading, writing, speaking, listening, viewing, and visually representing.

FIGURE 30 Descriptions of sections of student score report using a combination of text and graphics (taken from the interpretive guide included on the reverse side of a state's score report).

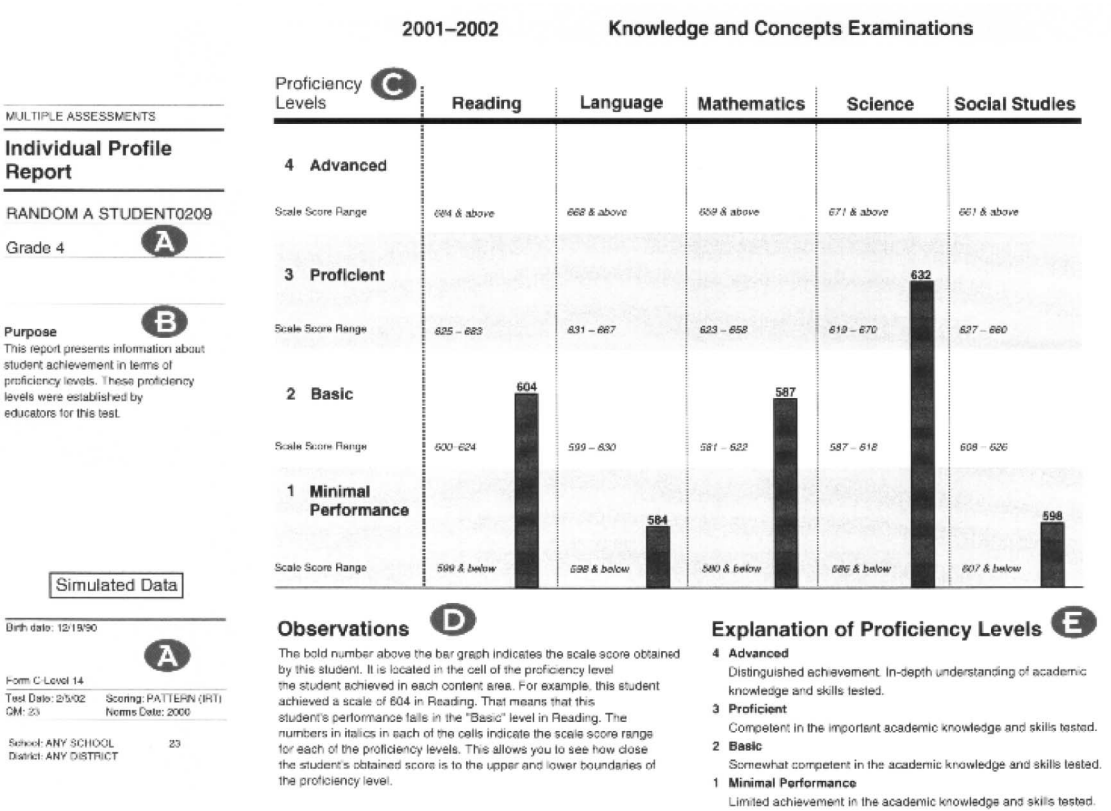


FIGURE 31 Description of sections of student score report using graphics and text (display from a state's interpretive guide). Reproduced from *Wisconsin Student Assessment System* (Wisconsin Department of Public Instruction, 2002). Copyright © 1997 by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc. Reproduced with permission. All rights reserved.

How Do I Read the Individual Profile Report?

A sample of the Individual Profile Report is found on page 5 of this guide.

The *Examinations* provide information about educational achievement and skills in basic content areas. The sample 4th-grade *Individual Profile Report* on page 5 indicates a 4th-grade student's proficiency levels in Reading, Language, Mathematics, Science, and Social Studies. The report will provide a complete record of the student's performance, including general information about achievements in these content areas as well as specific information about the student's levels of proficiency. This information can provide a basis for planning an educational program to meet specific academic needs.

Page 1: PROFICIENCY LEVELS

Ⓐ **Identifying Information.** This information can be found on the left-hand panel of the report. The student's name, grade, and birth date are shown. Also listed are the test Form/Level, test date, and norms date. School and district names and code numbers are listed in the lower part of this section.

QM 23: QM stands for Quarter Month. The report refers to QM 23, or the 23rd week of the school year.

Scoring: PATTERN IRT: Item Response Theory

Norms Date 2000: The most recent date a norms study was completed

Ⓑ **Purpose.** This statement indicates what the report contains and how to use the data. This helps teachers and parents/guardians interpret the test results.

Ⓒ **Proficiency Levels.** The chart provides information about student achievement in terms of proficiency levels. The proficiency levels are: Advanced, Proficient, Basic, and Minimal Performance.

For each subject, a scale score range for each proficiency level is indicated in *italics.* The bar graph indicates the proficiency level the student achieved in each content area. The **bold** number above the bar graph shows the scale score obtained by the student. This score determines the proficiency level the student attained. It must be within the scale score range of the proficiency level the student is shown to have attained.

For example, the sample report shows that this 4th-grade student achieved a scale score of 604 in Reading, as shown, in bold, above the bar graph. This score falls within the Reading scale score range of 600–624, the Basic level. This means that this student's performance falls into the "basic" proficiency level in Reading. This information shows how close the student's obtained score is to the upper and lower boundaries of the proficiency level.

Ⓓ **Observations.** This section provides individualized interpretive information about the student scores.

Ⓔ **Explanation of Proficiency Levels.** The lower right-hand portion of the page provides brief explanations of the proficiency levels. Refer to the proficiency descriptors beginning on page 7 of this guide for detailed descriptions of the proficiency levels for each content area.

FIGURE 31 *(continued)* Description of sections of student score report using graphics and text (display from a state's interpretive guide). Reproduced from *Wisconsin Student Assessment System* (Wisconsin Department of Public Instruction, 2002). Copyright © 1997 by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc. Reproduced with permission. All rights reserved.

UNDERSTANDING YOUR CHILD'S REPORT

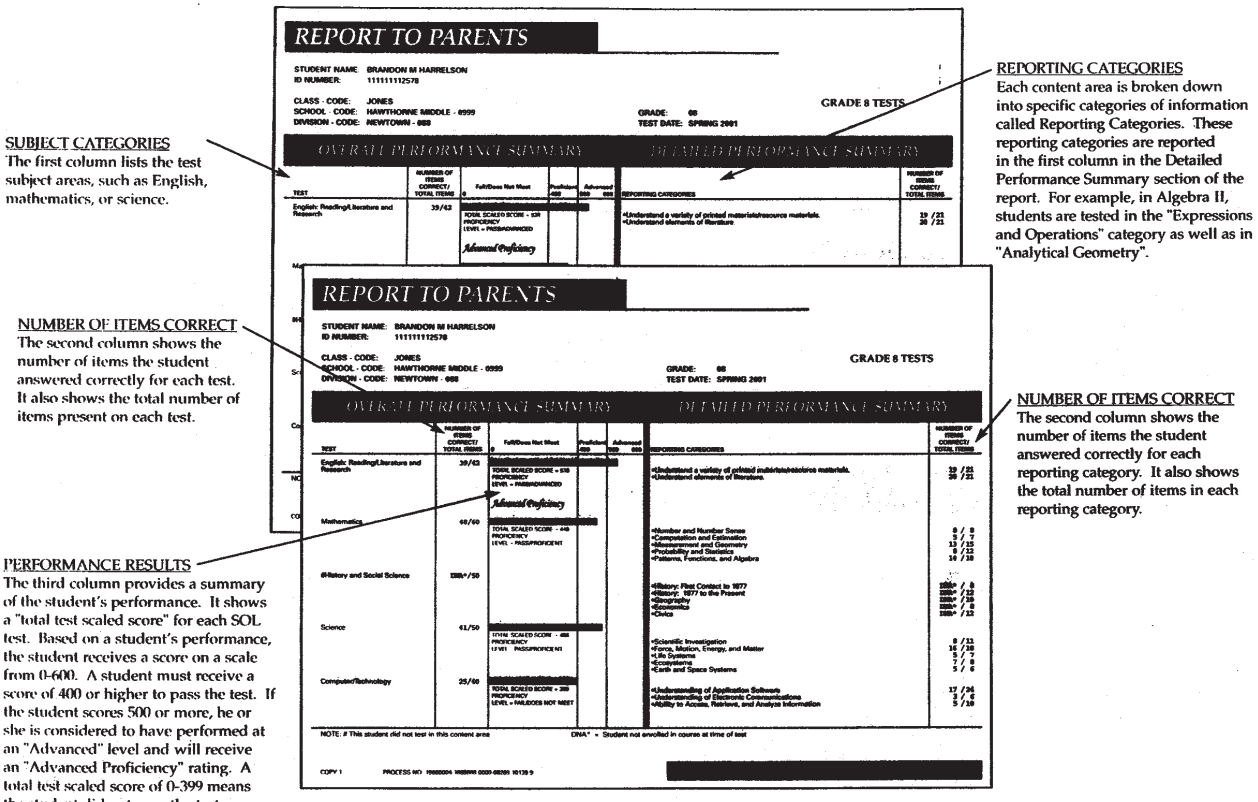


FIGURE 32 Description of sections of student score report using graphics and text (taken from the reverse side of a state's score report).

CURRENT STUDENT SCORE REPORTS AND INTERPRETIVE GUIDES: DISCUSSION OF FINDINGS

The review of recent student score reports and associated interpretive material produced by 11 states, 2 Canadian provinces, and 3 commercial test publishers illustrates (a) the many different types of information from large-scale assessments that are currently reported to parents, students, and teachers and (b) the various ways this information is reported. The variability in approaches is substantial and certainly raises questions about the level of success of these various approaches. Next, some promising and potentially problematic aspects of current reporting practices are discussed. The aim of this discussion is not to fault or praise individual states, Canadian provinces, or commercial test publishers for their reporting efforts. Instead, the intent is to outline some general and specific considerations and recommendations that may assist all states, Canadian provinces, and commercial test publishers in their continual efforts to improve student score reporting. In many instances, the best approaches will not be determined until more research is carried out via focus groups, "think-aloud" studies, and experimental research. In addition, the ideas discussed next should provide an important foundation for future empirical research on the efficacy of current and possibly more refined methods of reporting student-level assessment results. To increase the objectivity of the review of current reporting practices, the findings of this study are interpreted in relation to the results, recommendations, and requirements of the score-reporting literature and professional standards, as well as in relation to the key score-reporting requirements of NCLB.

Promising Features of Current Student Score Reports

Features That Appear to Make Reports More Readable

An important requirement of NCLB and professional and technical standards (AERA et al., 1999; NCME, 1995) is that student results be reported in an understandable manner. To be understandable, reports and their associated results must be readable. Although empirical evidence such as that collected by Hambleton and Slater (1997) and Wainer et al. (1999) is needed to determine the extent to which the student score reports are readable, several approaches appear to increase the readability of the reports, as described in the following sections.

Use of headings and other devices to organize reports. One particularly promising technique that makes the reports more readable is the use of large headings and other devices (e.g., boxes, lines, white space, and, to a lesser extent, color) to meaningfully organize the report into different components. Consistent with the recommendations suggested by Hambleton and Slater (1997), this technique is

used to some extent by all states, Canadian provinces, and commercial test publishers. Examples that illustrate the effective use of boxes and large headings are provided in Figures 1 and 30. Another promising technique is to phrase the headings in the form of key questions that will be answered by the adjoining information. Figures 7 and 9 illustrate how one state used this approach to introduce important pieces of information to parents. This suggestion has been made by H. Wainer (personal communication), among others.

Use of a highlight section. Another promising technique that helps make the reports more readable is the use of a distinct highlight section that provides readers with an overall summary of results. One such highlight section, which was reported in a box at the top of the report, is presented in Figure 5. The use of a highlights section appears to be a good approach for allowing parents to tell at a glance how their child performed (something that parents in the NEGP 1998 study stated was important).

Use of graphical displays. Although no single approach for reporting results is likely to be more effective in all respects than all other forms (Hambleton, 2002b; Tufte, 1983, 1990; Wainer, 1997b), the use of clear graphic displays appears to enhance the readability of the reports by drawing the reader's attention to major findings. Examples of two promising ways that graphic displays can be used to highlight results are provided in Figures 12 and 31. In both of these examples, the graphic displays allow readers to quickly determine how a student performed on different components of the assessment. An example of what appears to be a less effective graphic display (one that many teachers in the Impara et al. [1991] study had difficulty in interpreting) is shown in Figure 2. In this case, the graphical display does not provide a quick overview of how a student is performing in different subject areas (e.g., it is easy to confuse bars that represent overall scores with bars that represent subdomain scores). This display could probably be improved by labeling the bars or by using color or other devices (such as solid bars for overall scores and shaded boxes for subdomain scores) to help readers differentiate results from different subject areas and distinguish between overall and subdomain results.

Specially designed reports for different audiences. Another promising strategy for increasing the readability of student score reports is to create specially designed reports for different audiences. Recommended by Hambleton and Slater (1997), Jaeger (1998), and NEGP (1998) as a way to deal with the specific needs of different audiences, this strategy is used effectively by the two commercial test publishers that produced two versions of their student score reports. One version of the score report, targeted primarily to teachers, contains detailed diagnostic information about a student's performance on the assessment (see Figure 2). In contrast,

a parallel version targeted to parents (see Figure 3) contains substantially less data and would likely be easier to read. Although the creation of specially designed reports presumably will better meet the needs of a wide range of people, issues regarding access to all relevant information about a student's performance will still need to be addressed (e.g., when applicable, users of the simpler report should probably be informed that more detailed information about a student's performance is available from the student's school).

Personalized score reports. Several reports reviewed in this study embed the student's first name in multiple places throughout the report. This helps personalize the reports (something parents in the NEGP 1998 study appreciated) and appears to make the reports more inviting. Figures 1, 7, 11, and 16 illustrate how students' names are embedded in several reports. The successful application of this technique likely required procedures for obtaining an accurate first name for each student, as well as procedures for generating reports that can accommodate names of different lengths (e.g., either by providing a long blank space for inserting students' names within a statement [see Figure 7] or by the more sophisticated approach of making the surrounding text flush with the student's name [see Figures 1, 11, and 16]).

Features That Appear to Add Meaning for Intended Users of Student Score Reports

To further satisfy the requirement that results are reported in an understandable manner (NCLB, 2001; AERA et al., 1999; NCME, 1995), results also must be meaningful to intended users of student score reports. It is promising to note the many ways that states, Canadian provinces, and commercial test publishers are trying to make student results more meaningful to parents, teachers, and students. These include (a) describing the skills and knowledge assessed by the test (see Figure 1), (b) describing the expected levels of performance on the test through well-defined performance levels (see Figure 14), (c) describing the skills and knowledge a student possesses or does not yet possess (through use of performance levels or diagnostic information such as subdomain results [see Figure 17] and descriptions of specific strengths or weaknesses of particular students [see Figures 22 and 23]), and (d) reporting the results of relevant comparison groups (e.g., other students in the school, district, and state). These types of information will be very helpful in answering the key questions NEGP (1998, p. 36) recommends student reports answer (i.e., How did my child do? What types of skills or knowledge does his or her performance reflect? How did my child perform in comparison to other students in the school, district, state, and if available, the nation? What can I do to help my child improve?).

Another promising feature of the reports is that many results are reported in multiple ways (e.g., using numbers, graphics, and narrative text). Although perhaps increasing redundancies within reports, reporting assessment results in multiple ways should help address differences in the information processing needs and preferences of the many users of student score reports. Addressing these needs and preferences should help make the reports more meaningful to members of a diverse audience.

Figure 16 illustrates how one state reported results numerically, graphically, and narratively. Figure 11 shows how one commercial test publisher provided an easy-to-read narrative description of achievement scores displayed graphically and numerically on the score report (it also includes another promising feature: a blank space in which teacher comments can be written). These methods of reporting appear more promising than reporting results in a single way, such as numerically in a table (see Figure 4).

Reporting Results in Relation to Performance Levels

It is promising to note that 10 of the 11 states satisfy a key requirement of NCLB by reporting student results in relation to state performance levels. Both Canadian provinces also reported results in relation to provincial performance levels. Although not evident in the commercial test publishers' reports reviewed in this study, a study outlined in NEGP (1998) showed that at least two of the commercial test publishers produce score reports that display student results in relation to performance levels.

Different Ways to Report Results in Relation to Performance Levels

The findings of this study demonstrated that there are many different ways in which results are reported in relation to performance levels. The relative merits of each are discussed next.

A simple but effective way to communicate results in relation to performance levels is shown in Figure 12. In this example, the simple graphical display and accompanying text make it clear how a student performed in relation to the three performance levels. One novel feature of this report is that the results consider errors of measurement when classifying student performance according to the performance levels (e.g., the student's performance on the numeracy component falls somewhere between two performance levels). This did not appear to be the case with other reports that classified students into particular performance levels.

A potentially problematic feature of Figure 12 is that it does not indicate how close the student is to attaining a different performance level. This information, along with a graphic display of the precision associated with a student's score, is provided in the state score report displayed in Figure 7. Figure 16 also shows how

close a student is to attaining a different performance level, and provides a general description of the performance levels and two types of comparative information (averages for the district and state in relation to performance levels, percent of students in the district and state who achieved each performance level) that give additional meaning to a student's results.

Figure 13 illustrates how one state provides an overall summary of student performance in relation to performance levels, as well as providing more detailed information about the skills associated with the attained level of performance and how, on average, students in the state and district performed in relation to the standards. The overall summary provides a useful overview of how students performed across a number of subject areas and is consistent with Hambleton and Slater's (1997) recommendation that boxes and graphics be used to highlight main findings. The provision of skills associated with the attained level of performance and comparative information about how other students performed in relation to the standards are other promising strategies that NEGP (1998) suggested to enhance the meaning of the performance levels and a student's performance in relation to them. A potentially problematic feature of this display is that two performance standards (Level 2 and Level 3) are labeled in a manner that provides no insight into what the standards represent.

Figure 14 illustrates one particularly promising approach for reporting student results in relation to performance levels. As described earlier, this display includes a detailed description of the skills and knowledge represented by each performance level, providing clear insight into the types of skills and knowledge a student may need to develop to attain a higher level of proficiency. This display also uses a simple bar graph and easy-to-read labels that clearly highlight student performance. Figure 15 could be improved, however, through the use of bulleted lists, a larger font in the descriptions of the performance levels, and a darker bar graph for indicating student performance.

Reporting Diagnostic Information

Because states must report diagnostic information to satisfy the reporting requirements of NCLB, it is promising to note that all states and commercial test publishers include at least some diagnostic information in their student reports. This diagnostic information is reported in two general ways: as subdomain scores and as customized interpretations of the student's results.

Subdomain scores. Subdomain scores are the most common type of diagnostic information included in the reports reviewed in this study (8 of the 11 states, 1 of the 2 Canadian provinces, and all 3 commercial test publishers reported this type of information, typically as raw scores, percent correct scores, or percentile rank scores). The use of subdomain scores appears to be a promising way to

satisfy a key NCLB requirement and provides information about student's relative strengths and weaknesses that parents appear to value (NEGP, 1998). As will be noted, however, a number of problems may be associated with reporting this type of information, and care needs to be taken to ensure it is reported in an effective manner.

One way of reporting subdomain results is to provide the raw score obtained by a student in each content area (see, for example, Figure 17); however, raw scores have little or no meaning, and this practice is fraught with problems. In some reports, percent correct scores and graphical displays are provided (see, for example, Figure 18). But to really help provide meaning to these scores, it appears critical that general or detailed descriptions (see Figures 18 and 26, respectively) of the skills and knowledge that compose each subdomain are provided. Also, although possibly making displays more complex, providing baseline information at the subdomain level (such as the state average results provided in Figures 20 and 21) can help make students' subdomain scores more meaningful to users of the score reports. In one state, the meaning of student subdomain performance appears to be enhanced by providing the average subdomain performance of borderline proficient and borderline advanced students. Apparently, the average subdomain score of borderline proficient and advanced students is confusing to some users, but the state at least is grappling with the difficulty of trying to make subdomain scores more meaningful. Clearly, in all of these cases, research will be needed to judge added complexity against understandability and utility to intended audiences.

One particularly promising method of reporting subdomain results is presented in Figure 19. This display from a commercial test report includes two features not included in the score reports of any states or Canadian provinces: (a) an evaluation of student performance across subdomains in relation to specific levels of mastery and (b) information regarding the precision of the subdomain scores (reported graphically as confidence bands). The specification of ranges of scores that represent moderate levels of mastery across subdomain areas appears to be a promising way to further increase the meaning of subdomain scores. By providing information about the precision of all subdomain scores, Figure 19 satisfies an important professional and technical standard outlined by AERA et al. (1999). However, the provision of this information may be problematic in light of Hambleton and Slater's (1997) and Impara et al.'s (1991) findings that users of score reports have problems interpreting standard errors and percentile bands. Thus, there appears to be a clear need to educate users about how to interpret information about the precision of test scores whenever such information is provided.

Although subdomain scores appear to provide useful insight into the relative strengths and weaknesses of individual students in a given subject area and are being asked for by many teachers, parents, and students, two potential concerns about reporting these results should be noted. First, given the limited number of items involved in calculating subdomain scores, concerns may be raised about the

reliability of these scores and the validity of inferences drawn from them. One commercial test publisher's efforts to improve the accuracy of these scores through the combined use of item response theory and Bayesian estimation procedures appears to be a promising way to address these concerns. A second concern arises when subdomain scores are placed on a common scale (e.g., by reporting percent correct scores). Although facilitating comparisons across subdomains, placing these scores on a common scale may hide the fact that subdomain results may be based on different numbers of test items and item samples that are not equally representative of the relevant subdomains. To help address these concerns, the number of items that make up each subdomain score should be specified whenever these scores are reported, and users of the reports should be given clear guidance on how these results should be interpreted and used.

Customized interpretations of student results. A more sophisticated, but less widely used, approach to reporting diagnostic information is to provide written interpretations of an individual student's specific strengths and weaknesses on the score report (see Figures 22 and 23). Two states and one Canadian province used this approach to report diagnostic information.

These written interpretations of the test results appear to be a promising way to provide parents with a clear indication of their child's unique strengths and weaknesses. A potential advantage of this type of diagnostic information is that it does not require parents to derive meaning from numerous subdomain scores. It also appears to provide more specific information about student performance than is available through subdomain scores. Future research should investigate the relative advantages and disadvantages of using this approach to report diagnostic information versus reporting scores for relevant subdomain areas. This research also should explore the various issues involved in reporting customized interpretation of student results, such as the development of suitable interpretive statements and the manner in which they are assigned to students (e.g., through the use of individuals who score students' responses to open-ended items or through the use of computer programs that automatically identify statements based on students' subdomain scores).

Potentially Problematic Features of Current Student Score Reports

Although the student score reports reviewed in this study have many promising features, they also have potentially problematic features that warrant discussion.

Problematic Features Related to Reporting Results

Reporting too much information. Concerns raised by NEGP (1998) and the NRC (2001) about including too much information on assessment reports appear relevant to current reporting practice. Four reports reviewed in this study include numerous types of scores (e.g., two states report 3 different types of overall scores; two commercial score reports contain either 4 or 10 types of overall scores within a single subject area). This not only complicates the visual display of results but also increases the amount of technical jargon used in the report (something Hambleton and Slater, 1997, and the NEGP 1998 study have recommended minimizing). Figure 2 illustrates one commercial test report that may be problematic. In this instance, 9 different types of scores are reported across multiple domains and subdomains. Consistent with the recommendations of Impara et al. (1991), this score report would likely be less intimidating for teachers and parents if rarely used scores (e.g., NCE) were removed. Focus groups of intended users of the reports could be particularly helpful in identifying which scores should or should not be reported. A simpler display that would likely meet the needs of many users is shown in Figure 3.

Lack of information regarding the purpose of the assessment and how test results will be used. Although not widespread, one significant problem identified in this study is that not all of the reports outline the purposes of the assessment or explain how the assessment results will be used. Two states and one commercial test publisher appear to contravene professional and technical standards (AERA et al., 1999; NCME, 1995) by not including this important information on the student score reports. Because student score reports will be one of the primary sources of information that parents and students receive about many large-scale assessments, statements regarding the purpose of the assessment and how the assessment results will be used should be included on all of these reports. Inclusion of a purpose statement is particularly critical for new assessments that will be created as a result of NCLB, as well as for existing assessments whose purposes may change in response to this new legislation. One promising way two states and one commercial test publisher display their purpose statements is illustrated in Figure 31.

Lack of information regarding the precision of test scores. Another potentially problematic finding is that measures of precision are not regularly reported on student score reports. This is contrary to the professional standard that "score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level" (AERA et al., 1999, p. 148). Only four states and two commercial test publishers provide infor-

mation about the precision of overall test scores. In most cases, results reported in relation to performance levels do not include statements about measurement error. None of the states or Canadian provinces report information about the precision of subdomain results (although two states include general statements that indicate subdomain scores based on larger numbers of items are more reliable than subdomain scores based on smaller numbers of items).

Although the lack of clear statements of the degree of measurement error associated with each score is a violation of technical standards outlined by AERA et al. (1999), omitting such information does respond to the goal of keeping reports straightforward and clear. Hambleton and Slater (1997) and Impara et al. (1991) found that users of score reports often have trouble interpreting information such as standard errors and confidence bands. These conflicting recommendations help illustrate the difficulties associated with score reporting and the need for further research to determine how to best report technical information to a wide and diverse audience. What audiences are asking for may not be what they should receive if what they receive leads to misinterpretations of test scores.

The results of this study provide some insight into problematic and more promising approaches to reporting information about the precision of test scores. Figure 5 illustrates one problematic way of reporting the precision of test scores. In this example, standard errors associated with two scaled scores are reported without defining the term "standard error" or describing how standard errors should be used. Better ways of reporting the precision of test scores are illustrated in Figures 7 and 8. In these examples, probable ranges of scores are reported graphically. A numeric range is also provided in Figure 8; a description of what the range represents is included elsewhere on the report (i.e., if the student had taken the test numerous times, the scores would have fallen within the range shown).

Use of statistical jargon. Although not widespread, statistical jargon is present in some reports. Standard errors and NCE scores—data that posed a problem for participants involved in studies by Hambleton and Slater (1997) and Impara et al. (1991)—are reported in Figures 2 and 5. Percentile bands—data that most teachers could not interpret correctly in Impara et al.—are used in several reports (see Figures 2, 3, and 8). Scores that are not likely to be known even by many measurement specialists (e.g., Lexile measures) are reported with little or no information about what they mean or how they should be used.

Problematic Features Related to General Design

A number of problematic design features identified by NEGP (1998) and Hambleton and Slater (1997) are presented in many of the reports reviewed in this study. A small font is used in many reports, making text or numeric displays difficult to read (Figures 4 and 14 are examples of two displays that contain a small

font). Other potentially problematic features of the reports include the use of footnotes, abbreviations, and graphs that do not include scales.

At least three state reports reviewed in this study appear quite dense and cluttered, packing a lot of information in a limited amount of space. Several reports would likely be improved by a more judicious use of white space. Consideration should be given to either decreasing the amount of information reported or making the reports physically larger. For example, all state reports consist of one or two letter-sized pages. Six of these reports serve the dual roles of reporting a student's results and providing the only source of interpretive material for parents and guardians, all on a single letter-sized sheet of paper. Expanding the physical size of the reports may allow for a clearer design that could make better use of white space, use a larger font, and use other devices that may make the report more inviting to the user (e.g., an index, a separate glossary for key terms, content-relevant graphics). The folded 11″ × 19″ pamphlet used by one Canadian province and one commercial test publisher appears to be a very promising design for student score reports.

Promising Features of Current Interpretive Guides

Widespread Availability of Interpretive Material

It is promising to discover that interpretive material accompanies the student score reports of all participating states, Canadian provinces, and commercial test publishing companies. This finding is consistent with the recommendations of NEGP (1998) and Impara et al. (1991), and appears to satisfy the NCLB requirement that states produce "individual student interpretive…reports" (NCLB, 2001, §_1111[b][3][C][xii]).

Interpretive Guides Designed to Hold Score Reports

Two interpretive guides reviewed in this study are folders that are specially designed to hold student score reports. This folder design for interpretive guides is appealing because it (a) helps make the reports and guides appear to be part of a complete package and (b) allows for the communication of a large amount of interpretive information to the users of the score report. The commercial test publisher's interpretive folder with the inside pocket and resealable flap is impressive, although it will likely be too expensive for most states to produce on a large scale. The simpler folder produced by one state (created by folding a single 11″ × 17″ page in half) also appears effective and could likely be produced with relatively little expense.

Use of Graphic Displays to Describe Score Reports and to Provide Insight Into Test Questions

Several interpretive guides share the promising characteristics of reproducing the relevant score reports in the guide or including graphic displays of sample test

questions and of sample test questions in the interpretive guides. Graphic displays of the relevant score reports are useful in linking the various elements and sections of the score reports with relevant descriptions. Figures 30, 31, and 32 illustrate some promising ways to describe various sections of the score reports; these approaches appear to be much more effective than using only narrative text to describe the various sections of the score reports (see Figure 29). As recommended by NEGP (1998), sample test questions provide useful information about what the test measured, what it looked like, and what students should know and be able to do. They also appear to offer the added benefit of making the interpretive material more visually appealing. Figures 10 and 28 illustrate how sample test questions were included in the interpretive material produced by one state and one commercial test publisher. The use of graphical displays appears to be a promising way to make general descriptions of the test (e.g., Figure 27) more concrete and meaningful to the intended users of the score reports and interpretive guides. The potential value of using graphic displays in interpretive guides is a subject worthy of further research.

Attempts to Personalize the Interpretive Guides

A number of features appear to help personalize the interpretive guides. Several states include a letter to parents and guardians in their interpretive guide, often signed by a state official. These letters appear to help make the report more inviting and can help answer key questions parents have about the assessment (e.g., What is the purpose of the test? What was assessed? Where can parents get more information?) before they start reading and interpreting the results. Consistent with recommendations of NEGP (1998), a few interpretive guides include questions parents and students may wish to ask the teacher or school about the assessment results. Interpretive materials that accompany two score reports also include the promising feature of leaving a space where the teacher, parent, or student can write comments about the assessment. Pictures of students in the interpretive material of two states and two commercial test publishers also appear to help personalize the interpretive guides and assessment results.

Use of a Table of Contents

A promising feature of the interpretive guide produced by one state is the use of a table of contents. A table of contents helps to provide some order to the interpretive material and to facilitate retrieval of information. It is surprising that even some large documents (e.g., those that are 14 and 20 pages long) do not include a table of contents. States may wish to consider providing a brief table of contents in even relatively short interpretive guides (e.g., guides that are 4 pages or less). When the interpretive guides are relatively short, the table of contents might refer to numbered sections of the guide rather than entire pages. Future research might

explore the extent to which even a brief table of contents facilitates retrieval of information and accuracy of interpretations.

Availability of Interactive Web-Based Interpretive Guides That Complement Paper-Based Guides

Another promising finding of this study is the availability of interactive Web-based interpretive guides that complement information contained in paper-based guides. One state and one commercial test publisher produce these interactive Web-based guides, which use hyperlinks for easy retrieval of information.

For people who have access to the Internet, these electronic guides offer a rich source of information that can go well beyond what is traditionally available in a printed document (or PDF versions of a printed document). For example, the Web-based guides reviewed in this study provided more detailed explanations of key terms and concepts than the paper-based counterparts. These guides also may make test scores more meaningful by providing many different sample questions that are typically answered correctly by students with a particular score. Although this is a direction that states may wish to explore when reporting student results, states are cautioned to investigate and resolve potential concerns about access and equity that may arise from the use of interactive Web-based guides (e.g., equivalent print-based reports may need to be made available to those who do not have access to the Internet). Future research might also explore the affect Web-based guides have on score interpretation.

Potentially Problematic Features of Current Interpretive Guides

One potentially problematic physical characteristic of some interpretative guides is their length. At only 1 page in length, interpretive guides from seven of the participating states in the study may not provide users with sufficient information to accurately interpret the student test results. At 37 pages in length, one state's interpretive guide appears too long, providing excessive details that may impede the interpretation of the score report. Although including interpretive material on score reports may appear to be the most sound way to support proper interpretation of results, the physical constraints of most score reports reviewed in this study (i.e., reports that are either one or two pages long) limit the amount of interpretive material provided. Possible solutions include increasing the size of score reports to include more interpretive information or using folder-shaped interpretive guides designed to hold separate score reports. Future research should explore the extent to which separate interpretive guides versus interpretive material embedded in the score reports affect the interpretation of assessment results. Future research also should help identify an optimal length for the interpretive material.

Four of the interpretive guides reviewed in this study shared a second problematic feature of being overly dense. In all but one of these cases, too much material appeared to be included in a limited space (the 37-page interpretive guide simply overloaded the reader with too much information and text). These problematic guides do not make effective use of white space; typically do not organize the material using devices such as boxes, headings, or meaningful graphics; or use very small fonts. An example of interpretive material that appears to be overly dense is provided in Figure 29.

A third potentially problematic feature of the interpretive guides is that they do not always provide the types of information recommended in the score-reporting literature or required by technical and professional standards. For example, in apparent violation of professional standards (NCME, 1995), only about half of the states and Canadian provinces offer suggestions to help improve student performance.

A fourth potentially problematic feature of the interpretive guides is that key terms are not always defined. Although most key terms used in the student score reports are defined in the accompanying interpretive material, four states and one Canadian province do not define all key terms used in their score reports. In these instances, the most significant problem identified in this study is that performance levels are not always defined when they are used in reporting assessment results (3 of 10 states and 1 of the 2 Canadian provinces did not provide definitions for their performance levels in the student reports or interpretive material). This is especially problematic because students' achievement in relation to performance levels will be a critical consideration under NCLB.

For reports that include results for more than one subject area, one promising approach for providing definitions of performance levels is outlined in Figure 31. Here a brief explanation of performance levels is provided on the score report, with more detailed descriptions for each content area provided in the accompanying interpretive guide. For reports that include results from only one subject area, more detailed definitions of performance levels can be included in the student report (e.g., see Figure 14). In these cases, separate interpretive documents might include additional information such as sample questions and student responses indicative of the knowledge, skills, and abilities possessed by students at each performance level.

To help facilitate easy access to definitions, key terms defined in a special section of the interpretive guides and score reports (e.g., a glossary) would be helpful. Consistent with the findings of Hambleton and Slater (1997), technical terms should not be defined using footnotes.

CONCLUSIONS

In the next several years, individual and group score reports will be distributed in the United States at an unprecedented rate, providing student and statewide assessment results to millions of parents, students, teachers, educators, policymakers,

and members of the general public. Given the importance of reporting the information clearly and understandably, there is surprisingly little research to be found to guide the process of test score report design. What research is available appears to indicate that test score scales and reports are not always understood and used correctly. The intent of this study was to focus on student score reports only, to provide insight into current reporting practice, and to offer ideas that can facilitate improvement. Reports for aggregated results such as those required by NCLB legislation can be addressed in follow-up studies.

One purpose of this study was to identify sets of standards or recommendations that might exist for test score reporting. Several were identified and described in the first part of this article and should inform and guide the reporting efforts of states and publishing companies. The guidelines themselves are contained in the first three appendixes of this article for easy reference.

One of the main findings from the study is the tremendous variety of ways student-level assessment results are currently being reported. Whether good or bad, this variety is certain to be a rich source of ideas for states and test publishers as they work to improve their current methods of reporting and to meet the score-reporting requirements of NCLB and relevant professional standards.

By and large, it appears that current reports and interpretive guides are consistent with reporting standards but with some potential areas of weakness. Many states appear to be meeting key reporting requirements of NCLB, most notably reporting student results in relation to state performance levels and reporting some form of diagnostic information in student score reports. It was encouraging to find that score reports are accompanied by some form of interpretive material that provides meaning to the results and insight into the assessments; even more pleasing were efforts that help integrate interpretive material and score reports into a cohesive, informative package. States, Canadian provinces, and test publishers appear to be addressing the needs of many users of student score reports by reporting results in multiple ways, such as summarizing key results using easy-to-read narrative text, as well as using simple graphics that enhance data that are provided in numeric form. Approaches used by commercial test publishers also illustrate the value of creating alternate versions of student score reports that cater to the needs of different audiences (e.g., teachers vs. parents). Efforts to personalize the documents by embedding the student's name throughout the score report or by including informative letters to parents signed by a state official will be very much appreciated, as will be the use of illustrative graphics (e.g., sample test questions) that enhance the meaning of the results and make the reports and guides more visually appealing. This practice has been used successfully by the National Assessment Governing Board in reporting NAEP results, (see Hambleton & Smith, 1999).

Although some very promising features of current score reports and interpretive guides were identified through this study, a number of potential weaknesses warrant further attention. These include the following:

1. Excessive amounts of information (e.g., many types of overall scores) were included in some reports, and essential pieces of information (e.g., the purpose of the test, information about how the results will be and should be used) were not provided in others.

2. In many instances, information regarding the precision of test scores was not provided, making the results appear more accurate than they were.

3. Although not widespread, statistical jargon such as standard errors, NCE scores, and Lexile scores were present in more than a few reports.

4. Key terms, including the critical performance levels, were not always defined in the reports or interpretive guides, leaving the interpretations up to users, many of whom would be quite unaware of the proper interpretations to be made.

5. Efforts to report a large amount of information in a small physical space resulted in reports and interpretive guides that appeared dense and cluttered. Using small fonts was a common cause of concern across many reports and guides.

General Recommendations for Enhancing Future Student Score Reports

The review of current student score reports and interpretive guides, as well as their relevant strengths and weaknesses, suggested that many recommendations derived through research on NAEP or state-level reports or guidelines from literature on the visual display of quantitative information also are applicable to the creation of effective student score reports: (a) student score reports should be clear, concise, and visually attractive; (b) they also should include easy-to-read text that supports and improves the interpretation of charts and tables; (c) care should be taken to not try to do too much with a data display (i.e., displays should be designed to satisfy a small number of preestablished purposes); (d) devices such as boxes and graphics should be used to highlight main findings; (e) data should be grouped in meaningful ways; (f) small fonts, footnotes, and statistical jargon should be avoided; (g) key terms should be defined, preferably within a glossary (where they can easily be located by users); (h) reports should be piloted with members of the intended audience; and (i) consideration should be given to the creation of specially designed reports that cater to the particular needs of different users (e.g., a detailed score report may be appropriate for teachers but a simpler report may be more appropriate for widespread distribution to parents).

Seven additional recommendations can be derived from the findings of this score-reporting study and the identified strengths and weaknesses of current student score reports and interpretive guides. These are listed below, and, along with the previously stated recommendations, are summarized in Appendix D for easy reference.

First, include all information essential for proper interpretation of assessment results in student score reports. This includes a statement regarding the purposes of

the assessment, an explanation of how the results will and should be used, a description of relevant performance levels and test scores, and examples of how to interpret confidence bands. This would likely be best accomplished by increasing the size of student score reports from one double-sided page to two double-sided pages. A pamphlet design that is created by folding an 11″ × 17″ sheet of paper in half appears particularly promising. To facilitate photocopying and possible delivery of these reports in an electronic format (e.g., as PDF files available on a secure Web site), these score report pamphlets should be formatted in a manner that allows them to be easily printed onto four single-sided sheets (e.g., with blank margins and numbered pages). States will undoubtedly incur extra costs when printing these larger reports; however, if these costs are considered as being only a small percentage of the total printing budget for an assessment, it appears that the costs of producing a larger score report would be a sound and defensible investment (especially if they make the results of the assessment, as well as the assessment itself, more meaningful to such a large and diverse audience).

Second, include detailed information about the assessment and score results in a separate interpretive guide, ideally one in which the student score report can be inserted. A folded pamphlet design created from a double-sided 11″ × 17″ sheet of paper appears very promising (as with the pamphlet design recommended for the score report, states are encouraged to format this document in a manner that allows it to be easily printed on four letter-sized sheets). To facilitate retrieval of information, including a short table of contents in the interpretive guide would be helpful. Additional resources related to the assessment, such as relevant resource documents, Web sites, and telephone numbers of relevant contacts, also might be included in the interpretive guides for individuals who require further information.

Third, personalize the student score reports and interpretive guides. Embedding the student's first name strategically throughout the score reports appears to be one promising approach. To avoid inaccuracies, states interested in using this strategy are advised to establish procedures that will verify the accuracy of students' names before printing the student score reports. Another promising way to personalize the reports and interpretive guides is to explain basic information about the assessment in a letter to parents that is signed by a state official. Such personalized communications have been helpful in survey research and should be even more valuable with student test score reports. States might consider providing a space on the reports or guides in which teachers and parents can write comments about the student's results and questions or comments they may have about the assessment. More research, however, on this suggestion appears warranted.

Fourth, include an easy-to-read narrative summary of the student's results at the beginning of the student score report. This summary should highlight overall results, relevant diagnostic information, and pertinent implications of each.

Fifth, identify some things parents can do to help their child improve. Ideally, these suggestions would be included in a separate section near the end of the score

report and would be tailored to the level of performance demonstrated by the student in each subject area (e.g., those responsible for creating student score reports may wish to identify things parents can do to support learning for students with low, medium, or high performance on the assessment and include the relevant suggestions on a student's score report). Parents should be advised to talk with their child's teacher about other ways to improve performance. This was a very positive feature of several of the reports reviewed.

Sixth, include sample questions in the interpretive guides that illustrate the types of achievement represented by each performance level. The application of item response theory and item mapping procedures used in assessments such as NAEP would very useful in this regard (Zwick, Senturk, Wang, & Loomis, 2001).

Seventh, include a reproduction of student score reports in the interpretive guides to clearly explain the various elements of the reports.

Limitations of the Study and Suggestions for Follow-Up Research

This study represents only an initial attempt to gauge how student-level results on large-scale assessments are currently reported. It has a number of limitations, the most significant of which is the fact that it reflects the opinions of only two researchers. Although the views put forth in this article are informed by past score-reporting research and the comprehensive review of current reporting practices, it is clear that these views are not guaranteed to be right. The best way to begin to know what is right for designing score reports and interpretive guides is to generate a strong research base for designing this material.

In many instances, the best approaches for reporting student-level assessment results will not be known without the involvement of the intended users of these data. Future research should investigate the needs and preferences of parents, teachers, and students with respect to student score reports and interpretive guides. Interviews and focus groups are two particularly useful ways to obtain this type of information. Based on communications with participants in this study, it is clear that states, Canadian provinces, and commercial test publishers are continually refining their reports and reporting practices and are pilot testing proposed score reports with groups of people representative of the intended audiences. To help build a strong research base for designing score reports, test developers are encouraged to share promising and potentially problematic results of these operational activities with each other and the general measurement community.

In light of research that indicates preferences for a display and the readability of a display do not always go together (Wainer et al., 1999, p. 320), empirical studies should be conducted to determine the extent to which parents, teachers, and students understand student score reports. Think-aloud studies should provide valuable insight into the processes and assumptions that underlie the interpretation of

student assessment results (see, for example, Hambleton, 2002a). Promising and problematic features identified in this study could serve as an initial basis for experimental studies, and the recommendations outlined here should be evaluated using more objective procedures.

Given the importance of reporting diagnostic information on score reports, future research should explore ways to make this information more meaningful. The written interpretation of subdomain results used by two states in this study appears to be a promising way to present diagnostic information but should be evaluated empirically in relation to more traditional ways of reporting subdomain scores. States and researchers also should explore how to most effectively describe proficient performance on subdomain scores that are based on relatively few items.

Future research on reporting student-level results should involve a broad range of score reports. In this study, only two reports and typically one interpretive guide from each of the three commercial test publishers were reviewed. Some results of this study (e.g., results related to the use of performance levels in student-level reports) may have changed if all score reports produced by commercial test publishers were considered. In addition, the reporting practices of only 11 states and 2 Canadian provinces were considered in this study. Future research should explore how other states and Canadian provinces approach this important reporting activity. The findings of this study should be extended and validated by research that investigates how assessment results are reported for students in different grades. Studies that look at how results from multiple grades are reported will help determine the extent to which states satisfy the NCLB requirement that results are reported in a uniform manner. Building on current efforts to supplement traditional paper-based interpretive guides with Web-based versions, future research should explore the advantages and disadvantages of delivering score report material in an electronic format. Finally, research should investigate potential differences among members of different demographic groups with respect to the interpretation of assessment results and should identify ways to effectively communicate assessment results across these groups.

ACKNOWLEDGMENTS

This article is Center for Educational Assessment Research Report No. 477, Amherst, MA: University of Massachusetts, School of Education. It was presented as an invited paper at the meeting of the National Council on Measurement in Education, Chicago, IL, April 2003.

We wish to acknowledge the outstanding cooperation received from participating state test directors and test companies in the United States and provincial assessment representatives in Canada.

The first author gratefully acknowledges the financial support of Harcourt Assessment and the Harold Gulliksen Psychometric Research Fellowship Program at Educational Testing Service in completing this study.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

British Columbia Ministry of Education. (2002). *Foundation skills assessment: Individual student report.* Victoria, British Columbia, Canada: Author.

Connecticut State Board of Education. (2002a). *Connecticut academic performance test second generation: Interpretive guide.* Hartford: Author.

Connecticut State Board of Education. (2002b). *Connecticut academic performance test second generation: Understanding your teenager's scores.* Hartford: Author.

CTB/McGraw-Hill. (1997). *Missouri assessment program (MAP) student report.* Monterey, CA: Author.

CTB/McGraw-Hill. (2001a). *TerraNova, the second edition: A new concept of assessment information.* Monterey, CA: Author.

CTB/McGraw-Hill. (2001b). *TerraNova, the second edition: Home report and guide to the home report.* Monterey, CA: Author.

CTB/McGraw-Hill. (2001c). *TerraNova, the second edition: Individual profile report.* Monterey, CA: Author.

CTB/McGraw-Hill. (2003). *Interactive reports page.* Retrieved October 24, 2003, from http://www.ctb.com/mktg/terranova/tn_reports_noflash.jsp.

Data Recognition Corporation. (2003a). *2002 Pennsylvania system of school assessment grade 11 mathematics and reading: Individual student report.* Retrieved October 24, 2003, from http://www.datarecognitioncorp.com/paparent/grade5.html#grade11.

Data Recognition Corporation. (2003b). *2002 Pennsylvania system of school assessment grade 11 mathematics and reading: Individual student report.* Retrieved October 24, 2003, from http://www.datarecognitioncorp.com/paparent/main.html

Delaware Department of Education. (2002). *Delaware student testing program: A score results guide for parents.* Dover: Author.

Education Quality and Accountability Office. (2002a). *Individual student report: Ontario secondary school literacy tests, February 2002.* Toronto, Ontario, Canada: Author.

Education Quality and Accountability Office. (2002b). *Understanding the individual student report (ISR): Ontario secondary school literacy test (OSSLT), February 2002.* Toronto, Ontario, Canada: Author.

Education Week. (2003). *Quality counts 2003: If I can't learn from you.* Retrieved October 24, 2003, from http://www.edweek.org/sreports/qc03/.

Forte Fast, E., & the Accountability Systems and Reporting State Collaborative on Assessment and Student Standards. (2002). *A guide to effective accountability reporting.* Washington, DC: Council of Chief State School Officers.

Goertz, M. E., Duffy, M. C., & Carlson-LeFloch, K. (2001). *Assessment and accountability systems: 50 state profiles.* Retrieved October 24, 2003, from http://www.cpre.org/Publications/Publications_Accountability.htm.

Hambleton, R. K. (2002a, February). *A new challenge: Making results from large scale assessments understandable and useful.* An invited presentation at the Provincial Testing in Canadian Schools: Research, Policy, and Practice Conference, Victoria, British Columbia, Canada.

Hambleton, R. K. (2002b). How can we make NAEP and state test score reporting scales and reports more understandable? In R. W. Lissitz & W. D. Schafer (Eds.), *Assessment in educational reform* (pp. 192–205). Boston, MA: Allyn & Bacon.

Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

Hambleton, R. K., & Smith, T. (1999). *A focus group study of the general/public 1996 NAEP science reports* (Laboratory of Psychometric and Evaluative Research Report No. 361). Amherst: University of Massachusetts, School of Education.

Hamilton, L. S., & Koretz, D. M. (2002). Tests and their use in test-based accountability systems. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 13–49). Santa Monica, CA: RAND.

Harcourt Educational Measurement. (2002). *Stanford achievement test series.* 10th ed. Score report sampler. San Antonio, TX: Author.

Illinois State Board of Education. (2002). *No child left behind.* Retrieved October 24, 2003, from http://www.isbe.state.il.us/nclb/.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice, 10*(4), 16–18.

Jaeger, R. (1998). *Reporting the results of the National Assessment of Educational Progress* (NVS NAEP Validity Studies). Washington, DC: American Institutes for Research.

Joint Committee on Testing Practices (JCTP). (2004). *Code of fair testing practices in education.* Washington, DC: Author. Retrieved October 24, 2003, from http://www.apa.org/science/ jctpweb.html.

Landgraf, K. M. (2001). *Using assessments and accountability to raise student achievement.* Retrieved October 24, 2003, from ftp://ftp.ets.org/pub/corp/kurttest.pdf.

Linn, R. L. (1998). *Assessments and accountability* (CSE Technical Report 490). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

Louisiana Department of Education. (2002a). *2001–2002 interpretive guide: Grades 4, 8, 10, and 11 criterion-referenced tests.* Baton Rouge, LA: Author.

Louisiana Department of Education. (2002b). *Spring 2002 criterion-referenced test individual student report: English language arts—GEE21.* Baton Rouge, LA: Author.

Massachusetts Department of Education. (2002a). *MCAS tests of Spring 2002: Parent/guardian report.* Boston, MA: Author.

Massachusetts Department of Education. (2002b). *The Massachusetts comprehensive system: Guide to the MCAS for parents/guardians.* Boston, MA: Author.

Minnesota Department of Children, Families & Learning. (2002). *Minnesota comprehensive assessments: 2002 eleventh grade mathematics MCA student report.* Minneapolis: Author.

National Council on Measurement in Education (NCME). (1995). *Code of professional responsibilities in educational measurement.* Washington, DC: Author.

National Education Goals Panel (NEGP). (1998). *Talking about tests: An idea book for state leaders.* Washington, DC: U.S. Government Printing Office.

National Research Council (NRC). (2001). *NAEP reporting practices: Investigating district-level and market-basket reporting.* Washington, DC: National Academy Press.

New Jersey Department of Education. (2002). October 2002 high school proficiency assessment (HSPA): Cycle I score interpretation manual. Trenton: Author.

No Child Left Behind Act (NCLBA) of 2001, Pub. L. No. 107–110, §_1111, 115 Stat. 1449–1452 (2002).

Pennsylvania Department of Education. (2002). *Pennsylvania system of school assessment.* Harrisburg: Author.

Snodgrass, D., & Salzman, J. A. (2002, April). *Creating the Rosetta stone: Deciphering the language of accountability to improve student performance.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Taylor, A. R., & Tubianosa, T. S. (2001). *Student assessment in Canada: Improving the learning environment through effective evaluation.* Kelowna, British Columbia, Canada: Society for the Advancement of Excellence in Education.

Title I–Improving the academic achievement of the disadvantaged final rule, 67 Fed. Reg. 45,038 (Dec. 2, 2002) (to be codified at 34 CFR § 200.8[a]).
Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.
Tufte, E. R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.
Tukey, J. W. (1990). Data-based graphics: Visual display in the decades to come. *Statistical Science, 5*(3), 327–339.
University of Iowa. (2001a). *The Iowa tests: Interpretive guide for teachers and counselors.* Itasca, IL: Riverside.
University of Iowa. (2001b). *The Iowa tests: Report to students and parents.* Itasca, IL: Riverside.
Virginia Department of Education. (2002). *Virginia standards of learning assessments: Understanding your child's SOL report.* Richmond: Author.
Wainer, H. (1990). Graphical visions from William Playfair to John Tukey. *Statistical Science, 5*(3), 340–346.
Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher, 21*(1), 14–23.
Wainer, H. (1997a). Improving tabular displays: With NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics, 22*(1), 1–30.
Wainer, H. (1997b). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot.* New York: Copernicus Books.
Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36*(4), 301–335.
Wainer, H., & Thissen, D. (1981). Graphical data analysis. In M. R. Rosenweig & L. W. Porter (Eds.), *Annual Review of Psychology* (pp. 191–241). Palo Alto, CA: Annual Reviews.
Wisconsin Department of Public Instruction. (2002). *Wisconsin student assessment system: Interpretive guide for students and parents.* Madison: Author.
Wyoming Department of Education. (2002a). *The Wyoming comprehensive assessment system: Guide to interpreting the 2002 WyCAS reports.* Cheyenne: Author.
Wyoming Department of Education. (2002b). *WyCAS-Alt March 2002 student report for parents/guardians.* Cheyenne: Author.
Ysseldyke, J., & Nelson, J. R. (2002). Reporting results of student performance on large-scale assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students* (pp. 467–480). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20*(2), 15–25.

APPENDIX A
Standards from *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) Relevant to Reporting Student-Level Results on Large-Scale Assessments

- *Standard 5.10* When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used. (p. 65)

- *Standard 5.12* When group-level information is obtained by aggregating the results of partial tests taken by individuals (e.g., as is the case with matrix sampling), validity and reliability should be reported for the level of aggregation at which results are reported. Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established. (p. 65)
- *Standard 5.13* Transmission of individually identified test scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores. (p. 66)
- *Standard 5.15* When test data about a person are retained, both the test protocol and any written report should also be preserved in some form. Test users should adhere to the policies and record-keeping practice of their professional organizations. (p. 66)
- *Standard 5.16* Organizations that maintain test scores on individuals in data files or in an individual's records should develop a clear set of policy guidelines on the duration of retention of an individual's records, and on the availability, and use over time, of such data. (p. 66)
- *Standard 13.1* When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user. (p. 145)
- *Standard 13.7* In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision. (p. 146)
- *Standard 13.9* When tests scores are intended to be used as part of the process for making decisions for educational placement, promotion, or implementation of prescribed educational plans, empirical evidence documenting the relationship among particular test scores, the instructional programs, and desired student outcomes should be provided. When adequate empirical evidence is not available, users should be cautioned to weigh the test results accordingly in light of other relevant information about the student. (p. 147)
- *Standard 13.13* Those responsible for educational testing programs should ensure that the individuals who interpret the test results to make decisions within the school context are qualified to do so or are assisted by and consult with persons who are so qualified. (p. 148)
- *Standard 13.14* In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores. (p. 148)

APPENDIX B

Principles From the *Code of Fair Testing Practices in Education (Joint Committee on Fair Testing Practices, 2003a)*

Reporting and Interpreting Test Results	
Test Developers	*Test Users*
Test developers should report test results accurately and provide information to help test users interpret test results correctly.	Test users should report and interpret test results accurately and clearly.
C–1. Provide information to support recommended interpretations of the results, including the nature of the content, norms or comparison groups, and other technical evidence. Advise test users of the benefits and limitations of test results and their interpretation. Warn against assigning greater precision than is warranted.	C–1. Interpret the meaning of the test results, taking into account the nature of the content, norms or comparison groups, other technical evidence, and benefits and limitations of test results.
C–2. Provide guidance regarding the interpretations of results for tests administered with modifications. Inform test users of potential problems in interpreting test results when tests or test administration procedures are modified	C–2. Interpret test results from modified test or test administration procedures in view of the impact those modifications may have had on test results.
C–3. Specify appropriate uses of test results and warn test users of potential misuses.	C–3. Avoid using tests for purposes other than those recommended by the test developer unless there is evidence to support the intended use or interpretation.
C–4. When test developers set standards, provide the rationale, procedures, and evidence for setting performance standards or passing scores. Avoid using stigmatizing labels.	C–4. Review the procedures for setting performance standards or passing scores. Avoid using stigmatizing labels.
C–5. Encourage test users to base decisions about test takers on multiple sources of appropriate information, not on a single test score.	C–5. Avoid using a single test score as the sole determinant of decisions about test takers. Interpret test scores in conjunction with other information about individuals.
C–6. Provide information to enable test users to accurately interpret and report test results for groups of test takers, including information about who were and who were not included in the different groups being compared, and information about factors that might influence the interpretation of results.	C–6. State the intended interpretation and use of test results for groups of test takers. Avoid grouping test results for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use. Report procedures that were followed in determining who were and who were not included in the groups being compared and describe factors that might influence the interpretation of results.

C–7. Provide test results in a timely fashion and in a manner that is understood by the test taker

C–8. Provide guidance to test users about how to monitor the extent to which the test is fulfilling its intended purposes.

C–7. Communicate test results in a timely fashion and in a manner that is understood by the test taker.

C–8. Develop and implement procedures for monitoring test use, including consistency with the intended purposes of the test.

Informing Test Takers

Under some circumstances, test developers have direct communication with the test takers and/or control of the tests, testing process, and test results. In other circumstances the test users have these responsibilities.

Test developers or test users should inform test takers about the nature of the test, test taker rights and responsibilities, the appropriate use of scores, and procedures for resolving challenges to scores.

D–1. Inform test takers in advance of the test administration about the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Make such information available to all test takers.

D–2. When a test is optional, provide test takers or their parents/guardians with information to help them judge whether a test should be taken—including indications of any consequences that may result from not taking the test (e.g., not being eligible to compete for a particular scholarship)—and whether there is an available alternative to the test.

D–3. Provide test takers or their parents/guardians with information about rights test takers may have to obtain copies of tests and completed answer sheets, to retake tests, to have tests rescored, or to have scores declared invalid.

D–4. Provide test takers or their parents/guardians with information about responsibilities test takers have, such as being aware of the intended purpose and uses of the test, performing at capacity, following directions, and not disclosing test items or interfering with other test takers.

D–5. Inform test takers or their parents/guardians how long scores will be kept on file and indicate to whom, under what circumstances, and in what manner test scores and related information will or will not be released. Protect test scores from unauthorized release and access.

D–6. Describe procedures for investigating and resolving circumstances that might result in canceling or withholding scores, such as failure to adhere to specified testing procedures.

D–7. Describe procedures that test takers, parents/guardians, and other interested parties may use to obtain more information about the test, register complaints, and have problems resolved.

Note. From the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004, pp. 5-6). Washington, DC: Joint Committee on Testing Practices. (Mailing Address: Joint Committee on Testing Practices, Science Directorate, American Psychological Association, 750 First Street, NE, Washington, DC 20002–4242; http://www.apa.org/science/jctpweb.html.) Contact APA for additional copies. Adapted with permission.

APPENDIX C
NCME (1995) Professional Responsibilities of Those Who Interpret, Use, and Communicate Assessment Results

The interpretation, use, and communication of assessment results should promote valid inferences and minimize invalid ones. Persons who interpret, use, and communicate assessment results have a professional responsibility to:

6.1 Conduct these activities in an informed, objective, and fair manner within the context of the assessment's limitations and with an understanding of the potential consequences of use.
6.2 Provide to those who receive assessment results information about the assessment, its purposes, its limitations, and its uses necessary for the proper interpretation of the results.
6.3 Provide to those who receive score reports an understandable written description of all reported scores, including proper interpretations and likely misinterpretations.
6.4 Communicate to appropriate audiences the results of the assessment in an understandable and timely manner, including proper interpretations and likely misinterpretations.
6.5 Evaluate and communicate the adequacy and appropriateness of any norms or standards used in the interpretation of assessment results.
6.6 Inform parties involved in the assessment process how assessment results may affect them.
6.7 Use multiple sources and types of relevant information about persons or programs whenever possible in making educational decisions.
6.8 Avoid making, and actively discourage others from making, inaccurate reports, unsubstantiated claims, inappropriate interpretations, or otherwise false and misleading statements about assessment results.
6.9 Disclose to examinees and others whether and how long the results of the assessment will be kept on file, procedures for appeal and rescoring, rights examinees and others have to the assessment information, and how those rights may be exercised.
6.10 Report any apparent misuses of assessment information to those responsible for the assessment process.
6.11 Protect the rights to privacy of individuals and institutions involved in the assessment process.

APPENDIX D
Recommendations for Reporting Student-Level Results on Large-Scale Assessments

General Recommendations for Score Reporting

1. Score reports should be clear, concise, and visually attractive.
2. Score reports should include easy-to-read text that supports and improves the interpretation of charts and tables.
3. Care should be taken to not try to do too much with a data display (i.e., displays should be designed to satisfy a small number of preestablished purposes).
4. Devices such as boxes and graphics should be used to highlight main findings.
5. Data should be grouped in meaningful ways.
6. Small fonts, footnotes, and statistical jargon should be avoided.
7. Key terms should be defined, preferably within a glossary.
8. Reports should be piloted with members of the intended audience.
9. Consideration should be given to the creation of specially designed reports that cater to the particular needs of different users.

Summary of Recommendations for Reporting Student-Level Assessment Results

1. Include all information essential to proper interpretation of assessment results in student score reports (e.g., statements explaining the purpose of the assessment, the meaning of performance levels and test scores, and how the test results should be used, and examples of how to interpret confidence bands). Consider creating larger reports that can accommodate this information (a pamphlet design created by folding an 11″ × 17″ sheet folded in half appears particularly promising).

2. Include detailed information about the assessment and score results in a separate interpretive guide, ideally one in which the student score report can be inserted.

3. Personalize the student score reports and interpretive guides.

4. Include an easy-to-read narrative summary of the student's results at the beginning of the student score report.

5. Identify some things parents can do to help their child improve. Ideally, these suggestions would be included in a separate section near the end of the score report and would be tailored to the student's performance. Advise parents and guardians to talk with their child's teacher about other ways to improve performance.

6. Include sample questions in the interpretive guides that illustrate the types of achievement represented by each performance level.

7. Include a reproduction of student score reports in the interpretive guides to clearly explain the various elements of the reports.