Student Test Score Reports and Interpretive Guides: Review of Current Practices and Suggestions for Future Research [1, 2,3,]

Dean P. Goodman and Ronald K. Hambleton[4]

University of Massachusetts Amherst

June 19, 2003

---

[1] Center for Educational Assessment MCAS Validity Report No. 4 (CEA-477). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

[2] Invited paper presented at the meeting of the National Council on Measurement in Education, Chicago, April 2003.

[3] The research was completed as background for a contract between the Massachusetts Department of Education and the University of Massachusetts Amherst. However, the Department of Education is not responsible for any errors that remain in the paper nor should it be assumed that they agree with the findings and conclusions.

[4] The authors wish to acknowledge the outstanding co-operation received from participating state test directors and test companies in the United States, and provincial assessment representatives in Canada.

Introduction

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 towards 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, and used (Hambleton, 2002a). Concerns about the reporting of assessment results have been raised in recent years because there is a body of evidence showing confusion among policymakers, educators, and the public over the meaning and interpretation of large-scale assessment results (Hambleton, 2002a; 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 state 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 grade, will lead to unprecedented amounts of attention being directed toward state assessment results, especially at the individual student level.

At present, very little research 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 paper, student score reports and related interpretive guides from a sample of US states, Canadian provinces, and US commercial test publishers are reviewed. The purposes of this review are (a) to determine the types of information that are 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 first be considered.

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 et al. (2001)

showed that 45 states report student-level results on one or more statewide tests. As noted by *Education Week* (2003), more than half of these states are attaching significant stakes to these results (e.g., requiring students to obtain passing scores in order to graduate or to be promoted to 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 America is able to meet the high learning standards of the state where 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 three to eight 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] & § 1111[b][3][C][vii]). Beginning in 2007-2008, states will also 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, & National Council on Measurement in Education [NCME], 1999), *Code of*

Fair Testing Practices in Education (Joint Committee on Testing Practices [JCTP], 2003a), 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" (p. 2) 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). As many as 12 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 12 particularly relevant standards is provided in Appendix A.

The *Code of Fair Testing Practices in Education* (JCTP, 2003a) is a recent document 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. An updated version of the 1985 code, this draft resource is in the process of being endorsed by JCTP member associations and other organizations (JCTP, 2003b). A list of 14 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 (2002a), 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) provide clear evidence that many users of assessment data have difficulty interpreting and understanding results presented in large-scale assessment reports. While 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. While these reports have improved since about 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), and even when attempts were made at redesigning 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, 2002a; 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 making the report readable, concise, and visually attractive; keeping the presentation clear, simple, and uncluttered; not trying to do too much with a data display (i.e., displays should be designed to satisfy a small number of pre-established purposes); including text

to support and improve the interpretation of charts and tables; minimizing the use of statistical jargon; including a glossary of key terms; using bar charts to facilitate comparisons; grouping data in meaningful ways; using boxes or graphics to highlight main findings; avoiding the use of decimals; using color in a purposeful manner (given the potential for misuse, however, the general use of color was not universally recommended); piloting the reports with members of the intended audience; and 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 relating 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), being 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, they 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 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 (1998, 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 *prior* to the administration of the assessment, NEGP (1998, 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 comprising 11 parents from across the United States were also reported in 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. While 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 to be able to tell at a glance how their child performed. They also liked to see subtest scores and descriptions of the skills that were assessed by the test. Parents also appreciated learning what could 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 to what extent interpretive information provided on the reverse side of the student score report helped improve teacher understanding. While many teachers provided reasonable interpretations of information contained on the score reports, some types of information was misunderstood by large numbers of teachers. As noted by Impara et al. (1991), "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 information, most teachers (75% with interpretive information and 82.5% without such information) could not properly interpret percentile band performance profiles. Impara et al. (1991) 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. (1991) 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).

Current Student Score Reports and Interpretive Guides: Method and Results

Given 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 seemed warranted.

Method

The procedures for collecting and analyzing the reports and related material used in this study are outlined next.

Data Collection

Student reports and accompanying interpretive material were requested from the departments of education from 14 US states, three US commercial testing companies, and departments of education from two Canadian provinces. Reports and interpretive material for the tenth grade, or the grade closest to grade 10, were requested. Grade 10

was chosen since 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 was also obtained from three publishers of widely-administered commercial tests. These were Harcourt Educational Measurement (publishers of the *Stanford 10*), CTB/McGraw-Hill (publishers of the *TerraNova, The Second Edition*), and Riverside Publishing (publishers of the *Iowa Tests of Educational Development [ITED]*). British Columbia and Ontario, two of seven Canadian provinces that report student-level results on province-wide assessments (Taylor & Tubianosa, 2001), also submitted material for the study. It was hoped 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 only the home reports (Harcourt Educational Measurement, 2002, p. a; CTB/McGraw-Hill, 2001b, 2003; University of Iowa, 2001a). 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. This information 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.

4. Types of information included in the interpretive material and the manner in which they were reported. This information 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, provinces, and commercial test publishers are not identified when their reports are reviewed and discussed. As required, the names of the states, provinces, and commercial test publishers and any other obvious identifiers are removed from illustrative examples.

1. Contextual Features of Reports and Interpretive Guides

Grades and Year of Release

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

Stakes Assigned to Results

The stakes assigned by states and provinces to student-level results were identified from either 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, although without any explicit state-sanctioned stakes.

No stakes were attached to students' individual assessment results in one of the Canadian provinces, while high stakes (graduation from high school) were attached to the results in the other 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 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.

2. General Design Features

Physical Characteristics of the Student-Level Reports

The student-level reports from nine states consisted of two letter-sized 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 folded 11-inch by 17-inch pamphlet).

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-inch by 17-inch pamphlet that was 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 (less 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, 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 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 will release a separate interpretive guide in the near future.

The separate interpretive guides produced by four states were 4, 14, 20, and 37 pages in length. The four-page guide was printed on an 11-inch by 17-inch pamphlet that was folded in half (allowing the student report to be inserted inside the guide). The 14-page guide was 5.5 inches wide and 8.5 inches 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 inches wide and approximately 19 inches 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 would hold 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 of them made significant use of color). With a few exceptions, states and provinces did not use color in their interpretive guides.

Methods Used to Organize the Material

States, 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 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 has pioneered. Sections in all reports were also 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 in no other documents.

Insert Figure 1 about here.

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

Number of Subjects

Student reports from eight 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 two 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 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 province (six subjects on two commercial reports, and eight subjects on the third commercial report). All three commercial tests also reported composite scores, comprising 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 three 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 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 four or 10 types of overall scores (e.g., scale scores, grade equivalent scores, national and age stanines and percentiles, number correct, normal curve equivalent, Lexiles, 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 one (national percentile) or two (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.

Insert Figures 2 and 3 about here.

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 three 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 only reported overall scores numerically. The one 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 some other ways overall scores were reported by a state and a commercial test publisher.

Insert Figures 4, 5, and 6
about here.

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, and the one province that reported overall scores did not provide information about their precision. This information was reported graphically by two states and one commercial test publisher, numerically by one state, and both numerically and graphically by one state and one commercial test publisher. An example of how one 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.

Insert Figures 7 and 8
about here.

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 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 by neither 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 (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).

Insert Figures 9, 10, and 11
about here.

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 independent school district determining what standards were required for graduation). While 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 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 15 illustrate the ways three states and one province reported overall results in relation to performance levels.

Insert Figures 12 and 13 about here.

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

Insert Figures 14 and 15 about here.

Three states and one 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).

Insert Figure 16 about here.

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.

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 comprising 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 three commercial test publishers included some type of diagnostic information on their student reports (the home reports for two of the publishers did not include this information, however). One of the two 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 below.

Subdomain results. Eight of the 11 states, one of the two provinces, and all three 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 (e.g., numerically and graphically). Two examples of how subdomain results were reported are shown in Figures 17 and 18.

Insert Figures 17 and 18
about here.

No state or 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 were depicted in these commercial test reports.

Insert Figure 19 about here.

No state or 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 to a range of scores representing moderate mastery (see Figure 19).

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

In three states, student performance on each subdomain was compared to the average performance of all students in the state (see Figure 20), to the average performance of students in the state with the same reported history of instruction (see Figure 21), or, as described earlier, to 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.

Insert Figures 20 and 21 about here.

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 province. Figures 22 and 23 show how the two states reported particular strengths or weaknesses of individual students.

Insert Figures 22 and 23
about here.

4. 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 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.

Insert Figure 24 about here.

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

Insert Figure 25 about here.

Content assessed. All states, 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).

Insert Figure 26 about here.

What the test looked like. Eight of the 11 states, both provinces, and two of the three commercial test publishers provided information about what the test looked like (one of the seven 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 provinces, and two 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 one state described the types of items used in its assessment. Only two states and one commercial test publisher included sample questions in their interpretive guides (one of these states included sample questions only in its Web-based interpretive guide). Figure 28 illustrates how one commercial test publisher displayed sample questions in its interpretive guide. Two states identified separate resources that contained actual test items.

Insert Figures 27 and 28
about here.

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 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 provinces, and two of the three commercial test publishers provided recommendations about 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, 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 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 text and graphic displays (see Figures 30, 31, and 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 allowed users to get detailed information about different sections of the reports by clicking on the sections of interest.

Insert Figures 29, 30, 31, and 32
about here.

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

While commercial test publishers defined all key terms used in their student reports, four states and one province did not. Three states and one 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.

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, two provinces, and three 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 leads one to wonder about the level of success of these various approaches. Next, some promising and potentially problematic aspects of current reporting practice are discussed. The aim of this discussion is not to fault or praise individual states, 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, provinces, and commercial test publishers in their continual efforts to improve student score reporting. In many instances, the best approaches will not be known until more research is carried out via focus groups, think-aloud studies, and experimental research. Also, 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. In an effort to increase the objectivity of the review of current reporting practices, 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 key score reporting requirements of *NCLB*.

Promising Features of Current Student Score Reports

The student score reports from all states, provinces, and commercial test publishers have many promising features that will be described next.

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 are reported in an understandable manner. To be understandable, reports and their associated results must be readable. While 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 that appear to increase the readability of the reports are worth noting.

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, 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 key pieces of information to parents. This suggestion has been made by Wainer, 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 seems to be a good approach for allowing parents to tell at a glance how their child performed (something that parents in NEGP [1998] stated was important).

Use of graphical displays. While no single approach for reporting results is likely to be more effective in all respects than all other forms (Hambleton, 2002a; 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 graphic displays were 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 students are 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 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. While the creation of specially-designed reports will likely 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 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 seems 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 must also be meaningful to intended users of student score reports. It is promising to note the many ways that states, 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 15); (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). While 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 using narrative text. 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 where school or teacher comments can be written). These methods of reporting seem 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 ten of the 11 states satisfy a key requirement of *NCLB* by reporting student results in relation to state performance levels. Both provinces also reported results in relation to provincial performance levels. While 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 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 standards (Level 2 and Level 3) are labeled in a manner that provides no insight into what the standards represent.

Figure 15 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 and larger font in the descriptions of the performance levels and a darker bar graph for indicating student performance.

Reporting Diagnostic Information

Since 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 (eight of the 11 states, one of the two provinces, and all three 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 seem 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 simple, but seemingly effective, way of reporting subdomain results is to report the raw scores obtained by a student in each content area (see, for example, Figure 17). This type of information is augmented in some reports by percent correct scores and graphical displays (see, for example, Figure 18). To help provide meaning to these scores, it seems critical that general or detailed descriptions (see Figures 18 and 26, respectively) of the skills and knowledge comprising each subdomain are provided. While possibly making displays more complex, providing comparative information at the subdomain level (such as the state average results provided in Figures 20 and 21) may also help make students' subdomain scores more meaningful to users of the score reports. One state's provision of comparative subdomain results obtained by proficient students seems promising. 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 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 seems to be a clear need to educate users about how to interpret information about the precision of test scores whenever such information is provided.

While subdomain scores appear to provide useful insight into the relative strengths and weaknesses of individual students in a given subject area, 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 seemed to be one 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). While 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 comprise 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 report 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 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 should also 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

While 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

Four potentially problematic features related to reporting results are discussed next.

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 three 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 NEGP [1998] have recommended minimizing). Figure 2 illustrates one commercial test report that may be problematic. In this instance, nine 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. While 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. Since 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 information 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 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).

While 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.

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, with a description of what the range represents 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. While 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. (1991), 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., Lexiles) 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 present in many of the reports reviewed in this study. Small font is used in many reports, making text or numeric displays difficult to read

(Figures 4 and 15 are examples of two displays that contain 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 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, larger font, and other devices that may make the report more inviting to the user (e.g., an index, separate glossary for key terms, content-relevant graphics). The folded 11-inch by 17-inch pamphlet used by one province and one commercial test publisher seems to be a very promising design for student score reports.

Promising Features of Current Interpretive Guides

Six promising features of the interpretive guides are identified as a result of this study. These features are described next.

Widespread availability of interpretive material. It is promising to discover that interpretive material accompanies the student score reports of all participating states, provinces, and commercial test publishing companies. This finding is consistent with the recommendations of NEGP (1998) and Impara et al. (1991), and seems to satisfy the

NCLB requirement that states produce "individual student interpretive…reports" (NCLB, 2001, § 1111[b][3][C][xii]).

Interpretive guides that are 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 since 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 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-inch by 17-inch page in half) is also 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). 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 seems 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 seem to help personalize the interpretive guides and assessment results.

Use of 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 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 four 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 to what extent even brief table of contents facilitate 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 provide more detailed explanations of key terms and concepts than their 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. While 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 impact that 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 one page in length, interpretive guides from seven states may not provide users with sufficient information to accurately interpret the student test results. At 37 pages in length, one state's interpretive guide seems too long, providing excessive details that impede the interpretation of the score report. While 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 that is provided. Possible solutions include increasing the size of score reports to include more interpretive information, or using folder-shaped interpretive guides that are 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 impact the interpretation of assessment results. Future research should also help identify an optimal length for the interpretive material.

Four interpretive guides share a second problematic feature of being overly dense. In all but one of these cases, too much material is included in a limited space (the 37-page interpretive guide simply overloads the reader with too much information and text). These problematic guides do not make effective use of white space, and typically do not organize the material using devices such as boxes, headings, or meaningful graphics, or use very small font. 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 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. While most key terms used in the student score reports are defined in the accompanying interpretive material, four states and one 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 (three of ten states and one of the two provinces did not provide definitions for their performance levels in the student reports or interpretive material). This is especially problematic since 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 15). In these cases, separate interpretive documents might include additional information such as sample questions and student responses that are indicative of the knowledge, skills, and abilities possessed by students at each performance level.

To help facilitate easy access to definitions, key terms should be defined in a special section of the interpretive guides and score reports (e.g., a glossary). 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 astonishingly high rate, providing student and statewide assessment results to millions of parents, students, teachers, educators, policy-makers, 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, and what research is available seems 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 and 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 and/or recommendations that might exist for test score reporting. Several were identified and described in the first part of the paper and should inform and guide the reporting efforts of states and publishing companies. The guidelines themselves are contained in the three appendices of this paper 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 also 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, 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 versus 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, for example (see Hambleton & Smith, 1999).

While 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., multiple types of comparable 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 is not provided, making the results appear more accurate than they are.

3) While 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. Small font 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, suggests that many recommendations derived through research on NAEP or state-level reports or guidelines from literature on the visual display of quantitative information are also applicable to creation of effective student score reports: (a) student score reports should be clear, concise, and visually attractive, (b) they

should also 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 pre-established 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 font, 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 to 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-inch by 17-inch sheet of paper in half seems 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 seems 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-inch by 17-inch sheet of paper seems 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, might also 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 seems to be one promising approach. To avoid inaccuracies, states interested in employing 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 also consider providing a space on the reports or guides where 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 seems 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 also 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. While the views put forth in this paper 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.

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, 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, 2002b). 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 seems 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 should also 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. Additionally, the reporting practices of only 11 US states and two Canadian provinces were considered in this study. Future research should explore how other states and provinces approach this important reporting activity. The findings of this study should also 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 also 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.

References

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

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

Connecticut State Board of Education (2002a). *Connecticut Academic Performance Test Second Generation: Interpretive Guide.* CT: Connecticut State Board of Education.

Connecticut State Board of Education (2002b). *Connecticut Academic Performance Test Second Generation: Understanding your teenager's scores.* CT: Connecticut State Board of Education.

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

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

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

CTB/McGraw-Hill (2003). *Interactive reports page* [On-line]. Available: 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* [On-line]. Available at 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* [On-line]. Available at http://www.datarecognitioncorp.com/paparent/main.html

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

Education Quality and Accountability Office (2002a). *Individual student report: Ontario Secondary School Literacy Tests, February 2002.* ON: Education Quality and Accountability Office.

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

Education Week (2003). *Quality counts 2003: If I can't learn from you* [On-line]. Available: http://www.edweek.org/sreports/qc03/templates/ article.cfm?slug=17toc.h22

Goertz, M. E., Duffy, M. C., & Le Floch, K. C. (2001). *Assessment and Accountability Systems: 50 State Profiles* [On-line]. Available: http://www.cpre.org/Publications/Publications_Accountability.htm

Hambleton, R. K. (2002a). 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: Allyn & Bacon.

Hambleton, R. K. (2002b, 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.

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, MA: 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 (tenth edition) score report sampler.* San Antonio, TX: Harcourt Educational Measurement.

Illinois State Board of Education. (2002). *No child left behind* [On-line]. Available: 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 (2003a). *Code of fair testing practices in education (draft)* [On-line]. Available: http://www.apa.org/science/code-draft.html

Joint Committee on Testing Practices (2003b). *Draft projects of the Joint Committee on Testing Practices* [On-line]. Available: http://www.apa.org/science/jctpweb.html

Landgraf, K. M. (2001). *Using assessments and accountability to raise student achievement* [On-line]. Available: 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). *Spring 2002 criterion-referenced test individual student report: English language arts—GEE21.* LA: Louisiana Department of Education.

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

Massachusetts Department of Education (2002a). *MCAS Tests of Spring 2002: Parent/Guardian Report.* Boston, MA: Massachusetts Department of Education.

Massachusetts Department of Education (2002b). *The Massachusetts Comprehensive Assessment System: Guide to the MCAS for Parents/Guardians.* Boston, MA: Massachusetts Department of Education.

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

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

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

National Research Council (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. NJ: New Jersey Department of Education.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 1111, 115 Stat. 1449-1452 (2002)

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

Snodgrass, J. A., & Salzmann, D. (2002, April). *Creating the Rosetta stone: Deciphering the language of accountability to improve student performance.* Paper presented at the annual 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, BC: Society for the Advancement of Excellence in Education.

Title I–Improving the Academic Achievement of the Disadvantaged Final Rule, 67 Fed. Reg. 45038 (2002) (to be codified at 34 C. F. R. § 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: Report to students and parents.* Itasca, IL: Riverside Publishing.

University of Iowa (2001b). *The Iowa tests: Interpretive guide for teachers and counselors.* Itasca, IL: Riverside Publishing.

Virginia Department of Education (2002). *Virginia standards of learning assessments: Understanding your child's SOL report.* VA: Virginia Department of Education.

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.

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.

Wisconsin Department of Public Instruction (2002). *Wisconsin student assessment system: Interpretive guide for students and parents.* WI: Wisconsin Department of Public Instruction.

Wyoming Department of Education (2002a). *The Wyoming comprehensive assessment system: Guide to interpreting the 2002 WyCAS reports.* WY: Wyoming Department of Education.

Wyoming Department of Education (2002b). *WyCAS-Alt March 2002 student report for parents/guardians.* WY: Wyoming Department of Education.

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). Mahway, NJ: Lawrence Erlbaum Associates.

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, 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)

Standards 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 tests results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of the 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 13.5

When tests results substantially contribute to making decisions about student promotion or graduation, there should be evidence that the test adequately covers only the specific or generalized content and skills that students have had an opportunity to learn. (p. 146)

Standard 13.6

Students who must demonstrate mastery of certain skills or knowledge before being promoted or granted a diploma should have a reasonable number of opportunities to succeed on equivalent forms of the test or be provided with construct-equivalent testing alternatives of equal difficulty to demonstrate the skills or knowledge. (p. 146)

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*

C. Reporting and Interpreting 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-7. Communicate 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-8. Develop and implement procedures for monitoring test use, including consistency with the intended purposes of the test.

D. 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
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.

(Joint Committee on Fair Testing Practices, 2003)

Appendix C: NCME 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.

(NCME, 1995, Section 6)

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 pre-established 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 font, 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.

Summarized 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-inch by 17-inch sheet in half seems 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.

Figures

See separate files.

Insert before appendices.