



Research Memorandum

ETS RM–12-20

Applying Score Design Principles in the Design of Score Reports for *CBAL*™ Teachers

Diego Zapata-Rivera

Waverely VanWinkle

Rebecca Zwick

September 2012

ETS Research Memorandum Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Research Scientist

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Gary Ockey
Research Scientist

Donald Powers
Managing Principal Research Scientist

Frank Rijmen
Principal Research Scientist

John Sabatini
Managing Principal Research Scientist

Joel Tetreault
Managing Research Scientist

Matthias von Davier
Director, Research

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ruth Greenwood
Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Applying Score Design Principles in the Design of Score Reports for *CBAL*TM Teachers

Diego Zapata-Rivera, Waverly VanWinkle, and Rebecca Zwick
ETS, Princeton, New Jersey

September 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

<http://www.ets.org/research/contact.html>

Associate Editor: James Carlson

Reviewers: Irv Katz and Eric Hansen

Copyright © 2012 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and LISTENING. LEARNING. LEADING., are registered trademarks of Educational Testing Service (ETS). CBAL is a trademark of ETS.



Abstract

Every year in the United States, millions of score reports are produced and delivered to teachers, students, parents, school administrators, and policymakers. However, existing research shows that many educators have trouble understanding and making appropriate use of score reports. We have developed and evaluated several teacher score reports (i.e., individual, classroom, and item information score reports) in the context of a learning-centered assessment system, that is, the *CBAL*TM (Cognitively Based Assessments *of, for, and as* Learning) research project. In this paper we describe a framework for developing and evaluating score reports, present several Web-based teacher score report prototypes that have been designed by applying score design principles, and summarize the results of two studies designed to evaluate the score reports with teachers.

Key words: design and evaluation of score reports, score reports for teachers, design principles

Acknowledgments

We would like to acknowledge the members of the *CBAL*TM reporting team: Meg Vezzu, Janet Stumper, Debbie Pisacreta, and Tom Florek for their excellent work on the project. We would like to extend our gratitude to the teachers who participated in our studies. We also want to thank the members of the CBAL project for providing useful feedback and supporting this work. Finally, we thank Irv Katz and Eric Hansen for providing useful comments on an earlier version of this report.

Existing research on score reports indicates that teachers, school administrators, and policymakers have trouble understanding the terminology and graphical displays used to communicate assessment results (e.g., Hambleton & Slater, 1997; Lukin, Bandalos, Eckhout, & Mickelson, 2004; Zwick et al., 2008). Although principles for designing high-quality score reports have been proposed (e.g., Fast, 2002; Goodman & Hambleton, 2004; Hattie, 2009) and professional standards require test takers to be clearly informed about assessment results as well as the purpose and recommended uses of the assessment (e.g., American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999), many currently available score reports do not effectively convey this information to the intended audiences.

Our research focuses on designing and evaluating score reports that effectively communicate assessment information to the intended audiences. This work has been done in the context of ETS's Cognitively Based Assessment *of, for, and as* Learning (*CBAL*TM) research initiative (Bennett & Gitomer, 2009).

In this paper we describe a framework for designing and evaluating score reports and present various score reports for teachers that were designed and evaluated following the guidelines of the framework and particular score design principles. We also summarize the results of two studies designed to evaluate the score reports with teachers.

Related Research

Different streams of work deal with helping to create effective score reports for teachers. For example, heuristics for creating reports that communicate the intended message to a particular audience have been developed (e.g., Fast 2002; Goodman & Hambleton, 2004, Hambleton & Slater, 1997; Hattie, 2009; Underwood, Reshetar, & Leahy, 2006; Underwood, Zapata-Rivera, & VanWinkle, 2007). These heuristics build upon knowledge from related areas such as representing quantitative data using graphical representations (e.g., Tufte, 1983, 1996; Wainer, 1997, 2005) and designing graphical user interfaces (e.g., Nielsen, 1994).

Evidence also suggests that teachers require and would benefit from additional training on basic educational measurement concepts required to understand information that is usually included in score reports (Bennett & Shepherd 1982; Lukin et al., 2004; Zapata-Rivera, VanWinkle, & Zwick, 2010; Zwick et al., 2008).

Deng and Yoo (2009) collected an extensive list of useful score reporting resources that includes papers, guidelines, and sample score reports. These resources are available for researchers and practitioners who wish to know more about the state of the art.

Roberts and Gierl (2010) reviewed current score reporting practices, proposed a framework for developing score reports for cognitive diagnostics assessments, and showcased a score report for students in the domain of mathematics. The authors highlighted the importance of evaluating the score reports with the intended audience.

Relevant research also includes work on understanding cognitive load and its implications for communicating information effectively (Mayer, 2001, 2005; Mayer & Moreno, 2003; Sweller, 1999). Mayer (2001) presented a series of principles for developing effective multimedia applications based on empirical evidence.

Aware of existing research findings in the area, we have designed score reports for particular educational stakeholders (i.e., teachers, administrators, and students) that make use of external representations (e.g., graphs, text, tables, interactive multimedia components) to communicate assessment claims at different levels. For example, teacher reports may include task-level information, formative hypotheses (tentative claims that are subject to confirmation by other data sources), performance levels, and scaled scores based upon availability of supporting evidence (e.g., see the appendix). The next section describes a framework for designing and evaluating score reports.

A Framework for Designing and Evaluating Score Reports

This approach to designing and evaluating score reports is inspired by methodologies used in the following areas: assessment design (e.g., Mislevy, Steinberg, & Almond, 2003), software engineering (e.g., Pressman, 2005) and human-computer interaction (e.g., Nielsen, 1994). It includes the following steps: (a) gathering assessment information needs, (b) reconciling these needs with the available assessment information, (c) designing score report prototypes, and (d) evaluating these score report prototypes with internal and external experts. Figure 1 depicts this framework graphically.

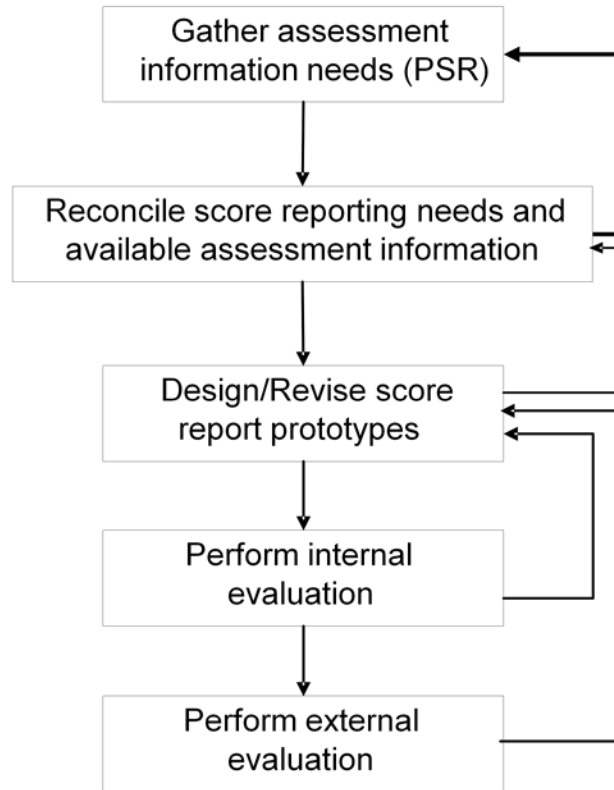


Figure 1. A framework for designing and evaluating score reports.

Gathering Assessment Information Needs

This phase involves gathering input about assessment information needs from various stakeholders including content experts and the intended audience(s). It may also include making use of information that has already been gathered, for example, results from prior assessment studies carried out with the same or a comparable audience. This information provides researchers with an initial view of what the users of the score reporting system expect. This information is captured in a document called the prospective score report (PSR) that is used to gather client requirements and serves as an input to the assessment development process. Information in the PSR is shared with content and measurement experts who can identify possible discrepant areas and provide appropriate suggestions for avoiding misunderstandings and unrealistic expectations that may result in disappointment for the user(s) of the score report.

Information for the PSR usually includes representations used on similar reports developed in the past (e.g., individual-, classroom-level reports), definitions of skills and subskills, possible performance levels, comparison, progress, and task-level information. The PSR provides a way for researchers to communicate their understanding of the reporting needs to content and measurement experts for their evaluation.

Reconciling User Needs With the Available Assessment Information

Any inconsistencies between what the intended audience expects and the internal assessment requirements need to be addressed during this phase. This generally implies making changes to the kind of assessment information that will be available (e.g., see the appendix) and how this information is presented in order to ensure that the intended audience receives the intended message, thereby reducing the likelihood of the audience making inappropriate use of report information.

As progress in the assessment development process is made, various score report prototypes are designed following best score report design practices. Each time changes are made to the requirements of the score report, the score report design needs to be updated (see cycle between “Reconcile score reporting needs and available assessment information” and “Design/Revise score report prototypes” in Figure 1). Also, any changes to the PSR based on outcomes from the assessment design process are shared with the client (see cycle between “Gather assessment information needs (PSR)” and “Reconcile score reporting needs and available assessment information” in Figure 1).

It is usually interesting to see how the PSR differs from the actual score report(s) produced by the assessment design process. These differences document the user initial expectations and its relation with the final product. Future work on the assessment (e.g., refinements) can be informed by the PSR.

Designing Alternative Score Report Prototypes

This phase involves designing score report prototypes that can be used to communicate the intended message to a particular audience. Best practices for designing high-quality representations should be followed, for example, work by Fast (2002), Goodman and Hambleton (2004), Hattie (2009), and Underwood et al. (2006). The use of preexisting score report templates that have previously been evaluated can facilitate this process. However, new elements

need to be designed to incorporate score report components that are not initially in the score report. Several score report variants are created to explore alternate representations. These variants may include different graphical representations, layouts, interpretive text, interactive components, and so on. Data to populate the score reports may be created to resemble actual data or may be actual data (if available). These score report designs are evaluated internally with the help of experts and externally with the intended audience(s).

Evaluating Score Report Prototypes Internally and Externally

Score report variants are evaluated first internally with the help of content, measurement, usability, and accessibility experts. Information gathered from experts is used to refine, create, or abandon score report variants. The resulting score reports are evaluated externally by conducting qualitative and quantitative studies with the intended audience. Acquired data are then used to refine the resulting score reports as well as to draw general lessons that can be used to improve the current state of the art in score reporting.

A similar framework for developing score reports is described in Hambleton and Zenisky (2010). This framework includes the following seven steps: (a) define purpose of score report, (b) identify intended audience(s), (c) review report examples/literature, (d) develop reports(s), (e) data collection/field test, (f) revise and redesign, and (g) ongoing maintenance. Roberts and Gierl (2010) described a framework for the development of diagnostic score reports that includes both content and form considerations. Roberts and Gierl's framework is based on research by Jaeger (1998) and Ryan (2003). All these frameworks seem to take a similar approach of taking into account the intended audience and the purpose of the assessment and then using current research finding and other resources to design score reports that are evaluated and refined in an iterative cycle. Our approach emphasizes the use of a prospective score report to keep in mind the requirements of the audience and describes the process of reconciling user expectations with the resulting report produced by applying an evidence-based assessment design methodology. In addition, our approach has been used to design and evaluate reports for various educational stakeholders including teachers, administrators, and students.

The next sections describe score reports created for teachers based on summative assessment results. Additional teacher reports from formative assessments are not included in this paper (Zapata-Rivera, 2011).

Score Reports for Teachers

Three types of score reports for teachers have been developed: individual, classroom, and item information. These score reports include traditional score report information (e.g., scaled scores, proficiency levels, and raw scores), interpretive text, a navigation pane, links to additional materials (e.g., skill definitions, sample tasks, and explanations of statistical terms used in the report), information about appropriate and inappropriate uses, and recommendations for teacher follow-up.

Figures 2 through 6 show a prototype of an individual student score report for teachers (mathematics). It includes five sections: introduction (Figure 2), appropriate and inappropriate uses (Figure 3), performance summary (Figure 4), task-level information on the current test (statistics and proportional reasoning; Figure 5), and a “What to Do Next” section with general recommendations for teacher follow-up based on student performance on the current as well as past tests (Figure 6). Additional materials such as general concepts and skill and task information are available through the vertical navigation pane as well as through the underlined hyperlinks integrated into the score report.

The screenshot displays the 'CBAL Individual Student Score Report' interface. At the top, a purple header bar contains the 'CBAL' logo and the title 'Individual Student Score Report'. To the right of the header, there are links for 'MATH Score Reports', 'Additional Materials' (General Concepts, Skill & Task Info), and navigation buttons for 'Back' and 'Next'. Below the header, a yellow bar provides student and test information: 'Student: Pat Smith', 'Teacher: Ms. Brown', 'Grade: 8', 'Test: 3 of 3', and 'Test Date: 03/30/10'. A vertical navigation pane on the left lists sections: 'MATH Score Reports' (with sub-links for Individual, Classroom, and Item Information), 'Introduction' (selected), 'Appropriate Use AND Inappropriate Use', 'Performance Summary', 'Current Test', 'What to Do Next', 'Additional Materials' (with sub-links for General Concepts and Skill and Task Information), and 'Help'. The main content area, titled 'Introduction', contains a letter to Ms. Brown explaining the report's purpose and structure. It lists the focus of each test: Test 1 (Model & Justify in Algebra), Test 2 (Justify & Develop Logical Argument in Proportional Reasoning), and Test 3 (Statistics and Proportional Reasoning). It also outlines the five main sections of the report: Personal Identification Section, Appropriate and Inappropriate Use Section, Performance Summary Section, Current Test Performance Section, and What to Do Next Section. At the bottom, a note states that the report includes links to additional information like skill descriptions and item information reports. The page is numbered '1 of 5'.

CBAL Individual Student Score Report

MATH Score Reports Additional Materials: General Concepts, Skill & Task Info BACK NEXT

Student: Pat Smith Teacher: Ms. Brown Grade: 8 Test: 3 of 3 Test Date: 03/30/10

Individual > Introduction >

Dear Ms. Brown,

Recently your students took the last periodic accountability assessment for the school year. The results presented here include information from the *current* test as well as information *aggregated* from past tests.

Each test has a focus (see below):

- Test 1: Model & Justify in Algebra
- Test 2: Justify & Develop Logical Argument (one-sided) in Proportional Reasoning
- **Test 3: Statistics and Proportional Reasoning**

The **Individual Student Score Report** is intended to provide you with individual student performance information on the current as well as past tests. This report can be used to help you identify student strengths and weaknesses in order to help inform your instructional decision making.

This score report has five main sections:

- **Personal Identification Section:** Includes student identification information, teacher name, grade, subject, test number and test date.
- **Appropriate and Inappropriate Use Section:** Describes the appropriate and some inappropriate uses of the test.
- **Performance Summary Section:** Includes the score, its confidence band, performance level, and distribution of scores information.
- **Current Test Performance Section:** Includes detailed information about student performance on the latest test (i.e., task-level performance information and sample tasks).
- **What to Do Next Section:** Summarizes student performance on the tests, presents information about what the student should do to progress to the next performance level, and offers general recommendations for teacher follow-up.

The score report includes [links](#) to additional information such as descriptions of skills and tasks, general measurement concepts and links to the Item information report.

1 of 5

Figure 2. Introduction (page 1 of 5).



Figure 3. Appropriate and inappropriate uses (page 2 of 5).

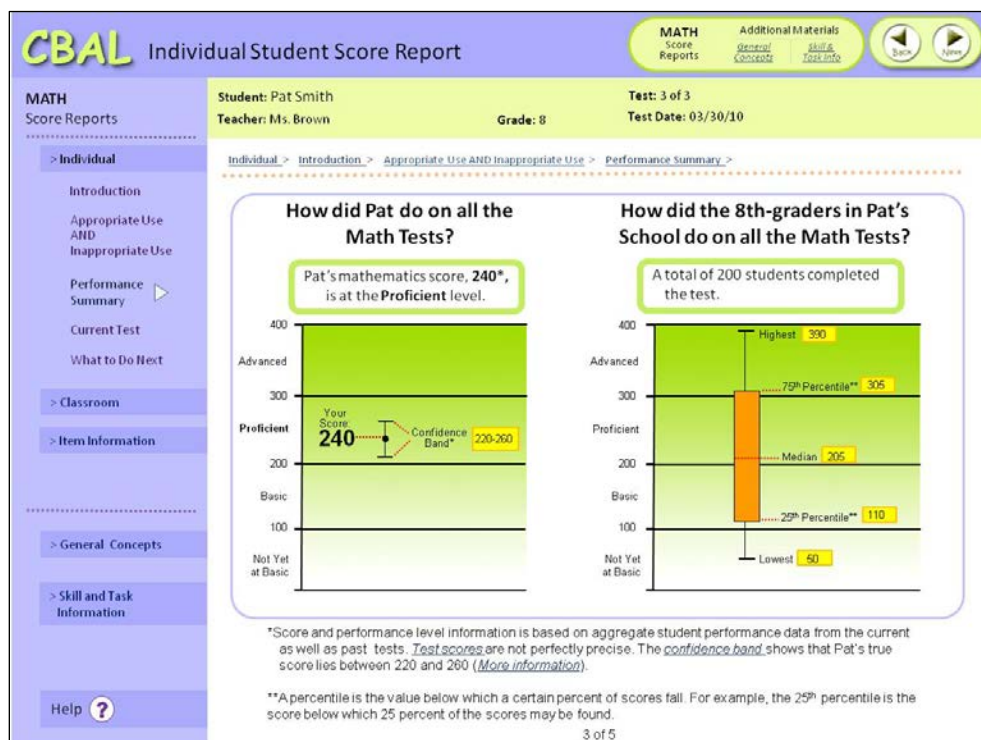


Figure 4. Performance summary (page 3 of 5).

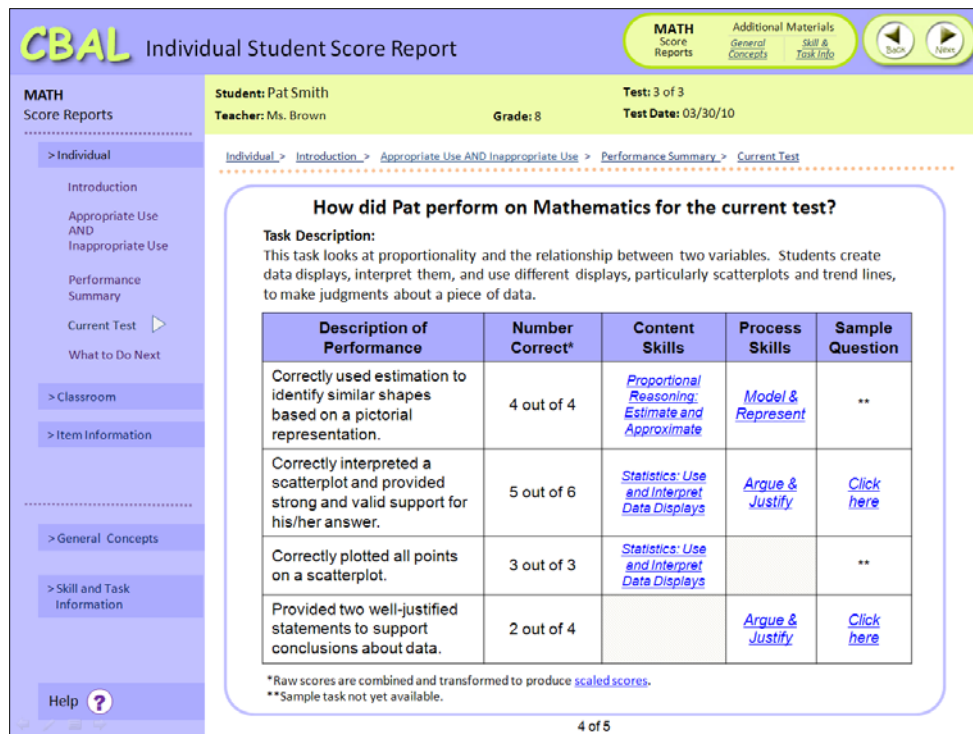


Figure 5. Current test performance (page 4 of 5).

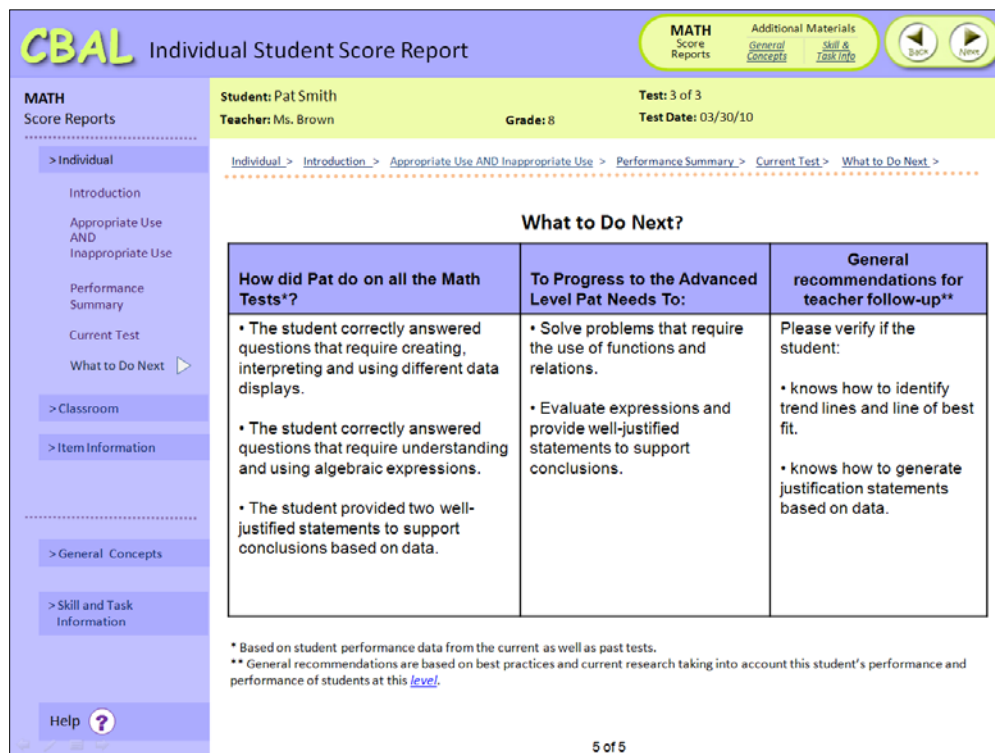


Figure 6. What to do next (page 5 of 5).

In addition to the individual student score report, two other score reports are available for teachers. First, a classroom score report includes the following sections: introduction (not shown), appropriate and inappropriate uses (not shown), and a sortable table showing classroom score and proficiency level information accompanied by an interactive graph depicting how the class is distributed among proficiency levels (see Figure 7). Individual student score reports can be accessed from the classroom score report by clicking on an individual's name.

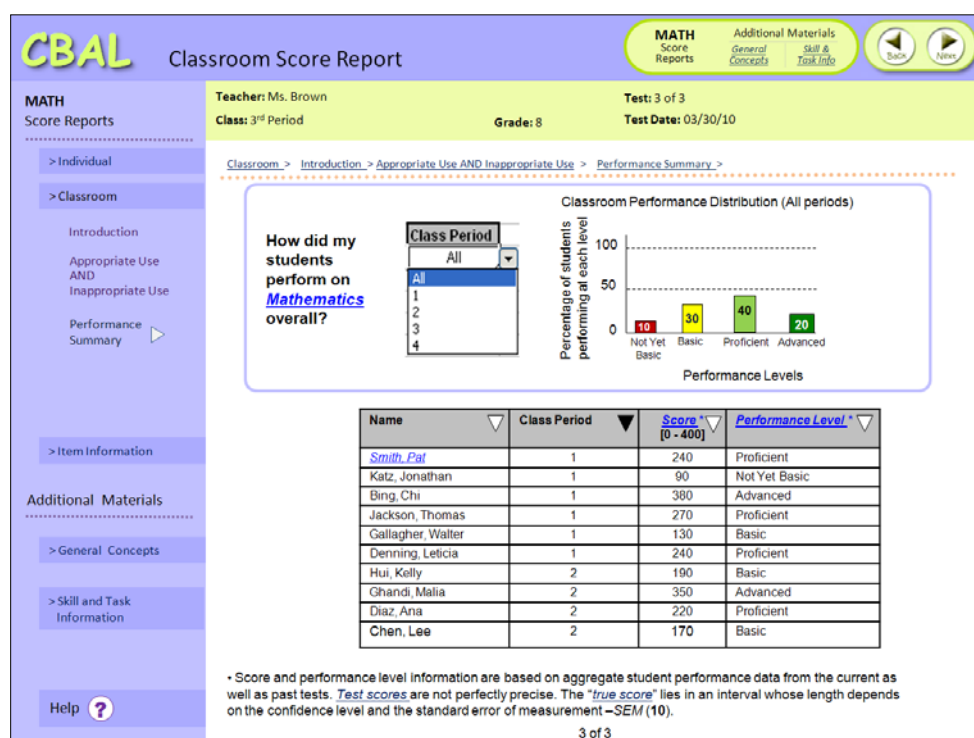


Figure 7. Classroom score report for teachers (math)—performance summary (page 3).
Names used in this report are not real student names.

Second, there is an item information score report that includes the following sections: introduction (not shown) appropriate and inappropriate uses (not shown), and the item difficulty table (Figure 8). Questions in this table are grouped by the content and process skills they share. Sample questions are also available through links.

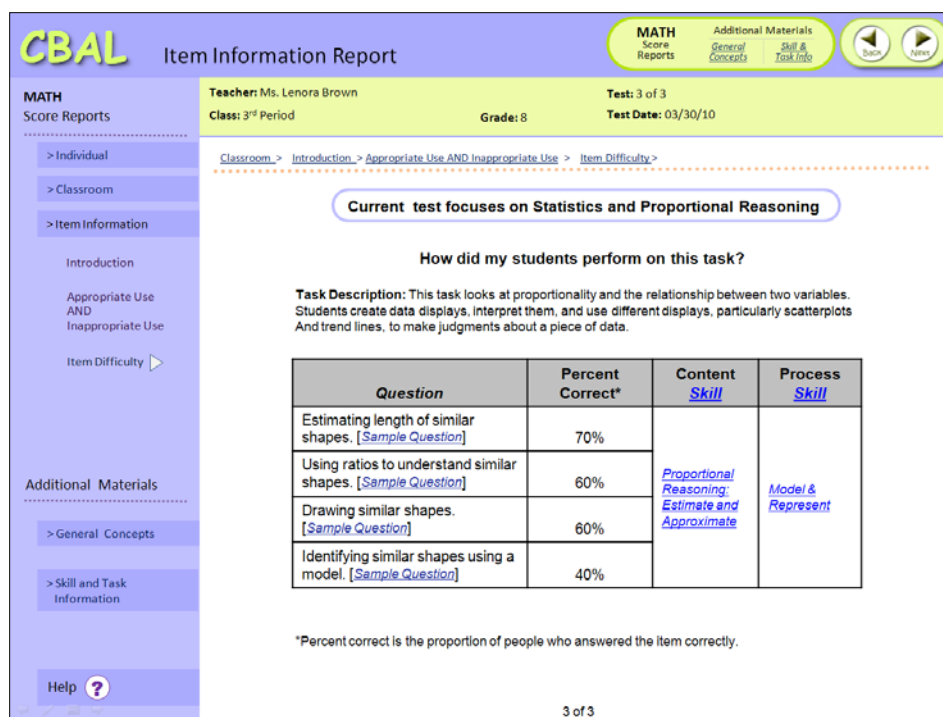


Figure 8. Item information report (math)–item difficulty (page 3).

Score Design Principles

Several score design principles were used to create the score report prototypes presented above. Some instances include the following:

- Present relevant information (Hattie, 2009; Mayer, 2001). Each page (section) of the score report responds to a particular question that is of interest to the teacher. Only the information needed to respond to the question is presented on the screen. Additional information is available through links (e.g., explanations and sample tasks). In Figure 4, two questions were included to facilitate comparison with the group.
- Minimize scrolling (Hattie, 2009). Instead of scrolling to each section of the report, teachers use the navigation pane on the left or the next and back buttons at the top of each screen.
- Use color in a purposeful way (Hattie, 2009; Tufte, 1983). This can be seen in the choice of colors for the interface components; the gradient green used to denote a continuous variable (i.e., scaled scores) and the color code used to show performance

levels in Figure 7 (i.e., red indicates there is a problem, yellow represents a warning, and green shows no problem—go).¹

- Order of presentation from general to detailed information (Goodman and Hambleton, 2004; Hattie, 2009). The sections of the report are organized following this principle. For example, performance summary information is presented first (i.e., scaled score and performance level; see Figure 4), followed by detailed information about the current test (task-level performance information; see Figure 5), and then information about what to do next (see Figure 6). Also, the classroom score report has a graph that shows the overall distribution of scores at the top of the screen followed by a sortable table that can be used to see detailed information.
- Use various representations (Mayer, 2001). Important summary information such as the scaled score and performance level of the student is presented both textually and graphically (see Figure 4).
- Place related words and pictures near each other on the page or screen (Mayer, 2001). This principle (of spatial contiguity) can be seen in the design of the graphs and tables used in the score reports. For example, Figure 4 shows information about the distribution of scores. This information includes a graphical representation, relevant concepts (labels), and values. In addition, information about these concepts can be accessed through links that are available in the explanation that appears below the graph.
- Provide information that helps users understand and make appropriate uses of score reports (Goodman & Hambleton, 2004; Hattie, 2009). The reports include information about appropriate and inappropriate uses, an introduction providing information about the report, and links to supplementary materials (definitions and sample tasks) that can be accessed through the links embedded in the report or through the navigation pane on the left of the screen.

Evaluation

Results of a usability study using a previous version of the score reports with 12 sixth-through eighth-grade teachers from schools in New Jersey and Pennsylvania showed that teachers reacted positively to interacting with the online score reports. Some of the teachers requested that hardcopy printouts be made available.

In general, the teachers preferred information presented as short, easy-to-read pieces. Long paragraphs were often ignored; after reading additional information (e.g., interpretive information and glossary entries), most teachers seemed to understand general concepts such as item difficulty, scaled scores, and raw scores. However, in general, teachers had problems understanding the concept of standard error of measurement. Most teachers identified and understood the purpose and appropriate use of each type of report. However, some teachers seemed to be willing to consider other uses and purposes that may or may not be appropriate (Zapata-Rivera & VanWinkle, 2010).

In a different study ($N = 147$), teachers were assigned to one of four conditions, which were obtained by crossing the two levels of report version (current vs. enhanced with additional links to relevant help topics for each page) with two tutorial conditions (tutorial administered versus tutorial not administered). After having interacted with the score reports (or a tutorial and the reports), participants were asked to complete a short comprehension test about information presented in the reports. Although the total test score variation was not statistically significant across experimental conditions, the responses to particular questions showed that teachers overall (across conditions) can recognize the correct use of each score report (94% or more). However, a significant proportion of the teachers were willing to consider other uses and purposes that were not valid (e.g., a valid use for the student score report is to place students in advanced or special programs, 54%; a valid use for the student score report is to evaluate the current math curriculum, 57%; a valid use for the student score report is to promote students to the next grade, 18%; and a valid purpose of student score report is to place students in advanced or special programs, 60%). The versions of the score reports employed in this study did not provide teachers with a list of inappropriate uses. In the current version of the reports, we have included both a list of appropriate uses and a list of inappropriate uses (e.g., see Figure 3). This study also showed that teachers overall had problems understanding key statistical concepts (e.g., reliability, 43%;

percentile, 54%; true score, 50%; and scaled scores, 42%). More information about this study can be found in Zapata-Rivera et al (2010).

Summary and Future Work

In this paper we present a framework for designing and evaluating score reports with particular audiences, describe score reports designed for teachers, illustrate how score design principles were used in the development of these reports, and present some results from two studies.

Future work includes refining the score reports (e.g., improving the wording of statistical information, minimizing the use of technical terms, and exploring alternative graphical representations), improving accessibility features, and conducting studies aimed at evaluating the use of various graphical representations and other report components to convey particular pieces of assessment information to this audience. In addition, we would like to explore how information gathered using these reports can be used to guide the development of formative materials for teachers and further explore using reports as communication tools to support the use of assessment information by teachers, students, parents, and other educational stakeholders.

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.
- Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Assessment issues of the 21st century* (pp. 43-61). New York, NY: Springer.
- Bennett, R. E., & Shepherd, M. J. (1982). Basic measurement proficiency of learning disability specialists. *Learning Disability Quarterly*, 5(2), 177–184.
- Deng, N., & Yoo, H. (2009). *Resources for reporting test scores: A bibliography for the assessment community*. Retrieved from https://ncme.org/resources/biblio1/NCME.Bibliography-5-6-09_score_reporting.pdf
- Fast, E. F. (2002). *A guide to effective accountability reporting: Designing public reports that effectively communicate accountability, assessment, and other quantitative education indicators in an easy understood format*. Washington, DC: Council of Chief State School Officers.
- Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, 17, 145–220.
- Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (Technical Report No. 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.
- Hambleton, R. K., & Zenisky, A. (2010, November). *Improvements to student score reporting: Steps and more use of suitable methodologies*. Invited presentation at the ETS Conference on Score Reporting, Princeton, NJ.
- Hattie, J. (2009). *Visibly learning from reports: The validity of score reports*. Retrieved from <http://www.oerj.org/View?action=viewPDF&paper=6>
- Jaeger, R. M. (1998). *Reporting the results of the National Assessment of Education Progress (NVS NAEP validity studies)*. Washington, DC: American Institutes for Research.

- Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, 23(2), 26–32.
- Mayer, R. E. (2001). *Multimedia learning*. Cambridge, England: Cambridge University Press.
- Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 31–48). New York, NY: Cambridge University Press.
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 43–52.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective*, 1(1), 3–62.
- Nielsen, J. (1994). *Usability engineering*. San Francisco, CA: Morgan Kaufmann.
- Pressman, S. (2005). *Software engineering: A practitioner's approach*. New York, NY: McGraw-Hill Education
- Roberts, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice*, 29(3), 25–38.
- Ryan, J. M. (2003). *An analysis of item mapping and test reporting strategies*. Retrieved from http://center.serve.org/_downloads/publications/Analysis_of_State_Test_Score_Reportin_g_Strategies.pdf
- Sweller, J. (1999). *Instructional design in technical areas*. Camberwell, Australia: ACER Press.
- Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
- Tufte, E. R. (1996). *Visual explanations*. Cheshire, CT: Graphics Press.
- Underwood J. S., Reshetar, R., & Leahy, S. (2006). *Score report design heuristics*. Unpublished manuscript.
- Underwood J. S., Zapata-Rivera D., & VanWinkle, W. (2007). *Growing pains: Teachers using and learning to use IDMS* (ETS Research Memorandum No. RM-08-07). Princeton, NJ: ETS.
- Wainer, H. (1997). *Visual revelations—Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. New York, NY: Copernicus Books.
- Wainer, H. (2005). *Graphic discovery*. Princeton, NJ: Princeton University Press.

- Zapata-Rivera, D. (2011, April). *Designing score reports that help teachers make instructional decisions*. Paper presented at the meeting of the Educational Research Association, New Orleans, LA.
- Zapata-Rivera D., & VanWinkle, W. (2010). *A research-based approach to designing and evaluating score reports for teachers* (ETS Research Memorandum RM-10-01). Princeton, NJ: ETS.
- Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2010, April/May). *Exploring effective communication and appropriate use of assessment results through teacher score reports*. Paper presented at the meeting of the National Council on Measurement in Education, Denver, CO.
- Zwick, R., Sklar, J., Wakefield, G., Hamilton, C., Norman, A., & Folsom, D. (2008). Instructional tools in educational measurement and statistics (ITEMS) for school personnel: Evaluation of three Web-based training modules. *Educational Measurement: Issues and Practice*, 27(2), 14–27.

Notes

- ¹ Future work includes improving accessibility features of the score reports (e.g., implementing cues that are redundant with color).

Appendix

Sample Claim Types and Evidence Requirements for Student-Level Reporting

	Type of claim			
	Task-level performance	Formative hypotheses	Performance level on total test and subscales	Location on a continuous scale for total test and subscales
Example	<p>There were 6 grammatical errors in this essay.</p> <p>Note: No claim is made about what the student knows or can do, only about the student's response to the test.</p>	<p>John may need to work on grammar, including subject-verb agreement.</p> <p>Note: The claim is tentative, subject to confirmation by other data sources available to the teacher (e.g., his or her own experiences with the student).</p>	<p><i>Total Test</i>: Meets Expectations <i>Formulate Arguments</i>: Meets Expectations <i>Assess Arguments</i>: Below Expectations</p> <p>Note: The claim is about what the student knows and can do.</p>	<p><i>Total Test</i>: 225 <i>Formulate Arguments</i>: 230 <i>Assess Arguments</i>: 175</p> <p>Note: The claim is about what the student knows and can do.</p>
Evidence requirement	Data supporting the accuracy of task-level performance characterizations (e.g., agreement with grammar error rates computed by a human judge)	Data indicating (a) agreement between the formative hypotheses from this test and those from another parallel test, (b) the consistency with which different raters generate formative hypotheses for a student from the same test responses, and (c) the relationship between the formative hypotheses and focused diagnostic measures or teacher judgments.	Data indicating (a) the probability a student's performance level from one set of tests would be the same as from another parallel set of tests, (b) the consistency with which tests are scored by different raters, and (c) the relationship between the student's performance classification and some independent classification measure (e.g., the current accountability test).	Data indicating (a) the relationship between scores on sets of parallel tests, (b) the consistency with which tests are scored by different raters, and (c) the relationship between the student's scale score and some independent measure (e.g., the current accountability test).
Availability	Across multiple occasions including the present one	Across multiple occasions including the present one	Across multiple occasions including the present one	For the last occasion, aggregated across all occasions