Listening. Learning. Leading.®

Research Report

ETS RR–11-45

Test Score Reporting: Perspectives From the ETS Score Reporting Conference

Edited by

Diego Zapata-Rivera

Rebecca Zwick

December 2011

Improving Test Score Reporting: Perspectives From the ETS Score Reporting Conference

Edited by

Diego Zapata-Rivera and Rebecca Zwick

Papers by

Jessica Hullman, Rebecca Rhodes, Fernando Rodriguez, and Priti Shah

University of Michigan, Ann Arbor

Rebecca Zwick

ETS, Princeton, New Jersey

Jeffrey C. Sklar

California State Polytechnic University, San Luis Obispo

Diego Zapata-Rivera

ETS, Princeton, New Jersey

December 2011

Technical Review Editors: Daniel Eignor and Donald Powers

Technical Reviewers: Richard Roberts, Diego Zapata-Rivera,
Lora Monfils, Carolyn Wentzel, Christine Lyon, Eric Hansen

Abstract

This volume includes 3 papers based on presentations at a workshop on communicating assessment information to particular audiences, held at Educational Testing Service (ETS) on November 4th, 2010, to explore some issues that influence score reports and new advances that contribute to the effectiveness of these reports. Jessica Hullman, Rebecca Rhodes, Fernando Rodriguez, and Priti Shah present the results of recent research on graph comprehension and data interpretation, especially the role of presentation format, the impact of prior quantitative literacy and domain knowledge, the trade-off between reducing cognitive load and increasing active processing of data, and the affective influence of graphical displays. Rebecca Zwick and Jeffrey Sklar present the results of the Instructional Tools in Educational Measurement and Statistics for School Personnel (ITEMS) project, funded by the National Science Foundation and conducted at the University of California, Santa Barbara to develop and evaluate 3 web-based instructional modules intended to help educators interpret test scores. Zwick and Sklar discuss the modules and the procedures used to evaluate their effectiveness. Diego Zapata-Rivera presents a new framework for designing and evaluating score reports, based on work on designing and evaluating score reports for particular audiences in the context of the CBAL (Cognitively Based Assessment of, for, and as Learning) project (Bennett & Gitomer, 2009), which has been applied in the development and evaluation of reports for various audiences including teachers, administrators and students.

Key words: graph comprehension, visual displays, graphs, visualization, score reporting, score interpretation, assessment literacy, teacher professional development, teacher education, score reporting to particular audiences, policymakers, administrators, teachers, students

Preface

Test results are used as evidence to support decision making at different levels of granularity. For example, decisions may pertain to individual students, classrooms, districts, or states. The information from tests needs to be understood and used correctly. Researchers in the area of score reporting have recognized the need for additional investment and work on designing and evaluating reports so that they clearly communicate assessment results to educational stakeholders. Advances in various disciplines, including educational measurement, cognitive science, human-computer interaction, and statistics can contribute to the development of innovative and effective score reports.

A workshop on communicating assessment information to particular audiences was held at Educational Testing Service (ETS) on November 4th, 2010. The goal of this workshop was to explore some of the issues that influence score reports and new advances that contribute to the effectiveness of these reports. The presenters were Ronald Hambleton (University of Massachusetts), Howard Wainer (National Board of Medical Examiners), Priti Shah (University of Michigan), Rebecca Zwick (ETS), and Diego Zapata (ETS).

This volume includes three papers that were written by presenters and their colleagues based on the presentations at the workshop:

- Jessica Hullman, Rebecca Rhodes, Fernando Rodriguez, and Priti Shah present results of recent research on graph comprehension and data interpretation. In particular, they consider the role of presentation format, the impact of prior quantitative literacy and domain knowledge, the trade-off between reducing cognitive load and increasing active processing of data, and the affective influence of graphical displays. They discuss the implications of these findings for the design of score reports for various audiences, including parents and educators.
- Rebecca Zwick and Jeffrey Sklar present the results of the Instructional Tools in Educational Measurement and Statistics for School Personnel (ITEMS) project, which was funded by the National Science Foundation and was conducted at the University of California, Santa Barbara. The goal of the project was to develop and evaluate three web-based instructional modules intended to help educators

interpret test scores. Zwick and Sklar discuss the materials that were developed and the procedures used to evaluate their effectiveness.

- Diego Zapata-Rivera presents work on designing and evaluating score reports for particular audiences carried out in the context of the CBAL (Cognitively Based Assessment of, for, and as Learning) project (Bennett & Gitomer, 2009). This work includes a new framework for designing and evaluating score reports that has been applied in the development and evaluation of reports for various audiences including teachers, administrators and students.

Papers by two of the presenters, Ronald Hambleton and Howard Wainer, are not included in this volume, but are available from the authors upon request (see also Wainer, 2009). Hambleton's presentation included examples of emerging methodologies for improving score report designs and evaluative criteria for use with student score reports. His findings were based on research conducted over the last ten years with the College Board, the National Assessment of Educational Progress, state departments of education, and several credentialing agencies. Wainer's presentation described the many factors that influence score report design. He suggested that the redesign of such reports should be guided by a sense of empathy with the examinee.

We would like to thank the presenters and internal reviewers. We are especially grateful to our editor, Ruth Greenwood, for her hard work and patience. Finally, we would like to acknowledge ETS for sponsoring this event and publication.

We hope the information in this volume informs the work of other researchers who wish to contribute to this area. We look forward to additional opportunities for collaboration.

Diego Zapata-Rivera and Rebecca Zwick

Table of Contents

Research on Graph Comprehension and Data Interpretation: Implications for Score Reporting

Jessica Hullman, Rebecca Rhodes, Fernando Rodriguez, and Priti Shah
University of Michigan, Ann Arbor, MI

Abstract

Score reports are frequently depicted in a graphic format. This chapter reviews some of the recent research in graph comprehension and data interpretation, and describes implications for score reporting. Specifically, the chapter discusses the research on the salience or impact of graphs and numbers, the influence of individual differences in graph comprehension, the possibility of tailoring information for different audiences, and a potential trade-off between ease of comprehension and desirable difficulties that encourage individuals to process information more deeply.

Key words: graph comprehension, visual displays, graphs, visualization, score reporting

Scores from educational tests are reported to a variety of audiences, including researchers, administrators, policy makers, politicians, teachers, parents, and students, often for different reasons. In almost all cases, scores are reported using numbers and graphs. In this chapter, we discuss current psychological research on graph comprehension and data interpretation as they relate to score reporting. We note here that we do not provide a comprehensive review of graph comprehension or data interpretation (for a relatively recent comprehensive review of graph comprehension, see Shah, Freedman, & Vekiri, 2005). Rather, we focus on current findings that we suggest may have implications for score reporting.

In the first section of the paper, we describe some basic psychological findings about the effect of different graphic formats on the comprehension of quantitative data. This section highlights research that identifies the most likely, salient interpretation of data, given a particular format. The research on graph comprehension provides the foundation for the next three sections of the paper that address three possible concerns for individuals who design score reports. The first concern is the potential over-reporting of underspecified or unreliable constructs (Twing, 2008). Contributing greatly to this concern is the fact that information presented in graphs is highly salient and may even lead to greater affective responses than information presented numerically. Consequently, graphically presenting score information that is not reliable or well-defined, such as a subscore that relies on few observations, may lead to an overuse of those numbers and perhaps even more positive or negative reactions than warranted. Furthermore, making some information more visually salient than other information may lead to additional interpretation errors.

The second concern for score report designers regards the different goals and abilities of the audience. Although one individual difference—statistical and quantitative literacy—is frequently the focus of investigation with respect to individuals' understanding of graphs and data (Shah et al., 2005), we argue here that individual differences in prior knowledge and dispositions can also have an impact on the interpretation of score reports. We discuss research on tailoring of graphical displays for different audiences and ability groups, and we suggest how this research might be applied to score reporting.

The third concern is that individuals may not critically evaluate information presented, but instead focus on one or two salient bits of information. Consider, for example, a score report that graphically depicts a large reduction in the achievement gap (Figure 1).

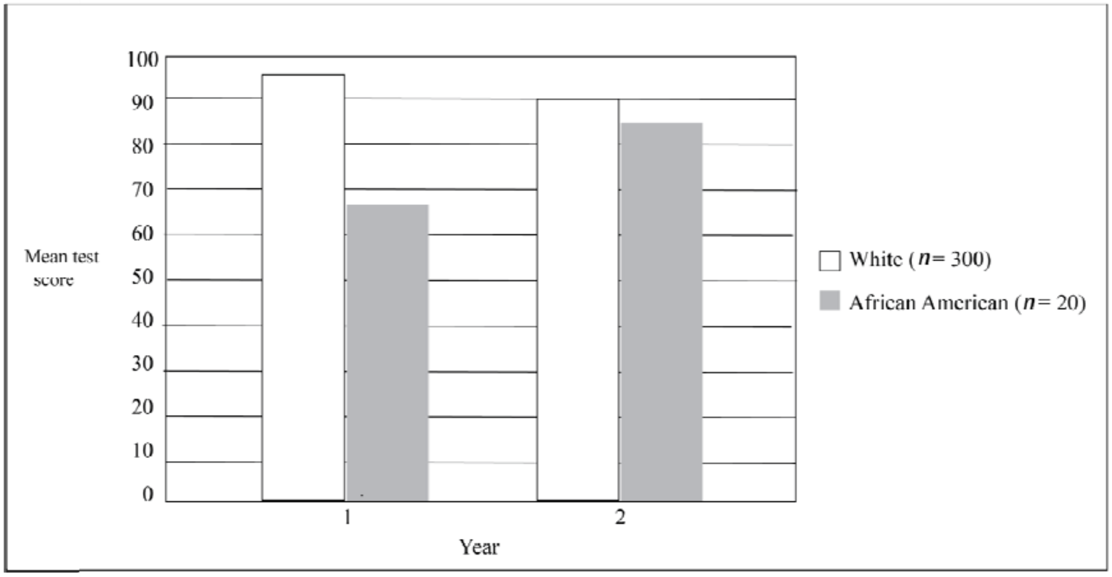


Figure 1. **Mock display representing achievement gap in test scores across two years.**

Research on graph comprehension would predict that the most likely interpretation when viewing multiple pairs or clusters of bars is to focus on the relative difference between the two bars on the left and compare that difference to the two bars on the right (Shah & Freedman, 2009; Shah, Mayer, & Hegarty, 1999). Thus, the most likely interpretation of the data is that there is a large difference between White and African American students in Year 1, but that this difference has virtually disappeared in Year 2. If a school administrator viewing these data focuses on the most salient information, he or she may not step back and think more deeply about the information presented and, in particular, the number of individuals that each bar represents. In reality, what appears to be a large reduction in the achievement gap is actually driven by a very small number of scores. When some information is made too readily accessible, we argue, people might form biased or oversimplified interpretations; in such cases, a more complex display or multiple graphic formats may be needed.

In this paper, then, we first describe the effect of graph format on viewers' interpretations of data. Next, we discuss how different variables—such as the salience of graphs and numbers; individual differences in not only numeracy and graphical literacy, but also differences in dispositions and prior content knowledge; and displays in which some interpretation readily "pops out"—can all lead to potential problems in the interpretation of score reports. Finally, we provide guidelines to avoid these possible interpretation errors.

Effect of Format on Graph Comprehension

Cognitive models of graph comprehension suggest that visual elements (e.g., the symbols, colors, types of lines, shape fills) are encoded, identified, and grouped together into chunks (Pinker, 1990). These "visual chunks" influence viewers' interpretations of the data. As discussed below, bottom-up factors, such as format (line or bar graph), influence the nature of those visual chunks. Specifically, a display is chunked based on the Gestalt principles of proximity, good continuity, and similarity (Pinker, 1990). Viewers map the salient visual chunks onto quantitative relationships or facts and then relate the quantitative information to meaningful referents. For example, in Figure 1, the large difference in the two bars on the left is often noted. The viewer must associate one bar with White students and the other to African American students and associate the height of the bars with test scores.

Much research on graph comprehension has focused on the relative ease and accuracy of retrieving or making inferences about data depicted in different ways, primarily because format affects the salience of particular visual chunks. Thus, the same data, depicted in different graphic formats, can have a large influence on viewers' interpretations of the data. A review of this body of research is beyond the scope of this chapter (but see Shah et al., 2005, for a review). In this chapter, we outline several basic findings regarding viewers' interpretations of some basic graphic formats, including tables, bar graphs, line graphs, and pie charts.

When viewing tables, viewers are often able to accurately encode individual data points, but have difficulty making inferences about trends (e.g., Guthrie, Weber, & Kimmerly, 1993). For example, if an administrator is viewing a table of scores on different subtests across different years, he or she might be able to compare different pairs of individual scores very well (i.e., in 2009, math scores were higher than in 2008). However, he or she may have difficulty noting that scores were increasing more rapidly for several years and that changes were leveling off, or that the relative improvements in math scores were in contrast to relative declines in reading scores. Despite the fact that tables make it difficult to read trends, it is not necessarily a good idea to avoid tables in all circumstances. In fact, tables can be beneficial for comprehension because they are relatively "neutral" to interpretation, unlike graphs, which can frequently bias individuals' interpretations. In other words, a reader might have to work harder to get the information wanted from a table, but the initial format would not have an influence. At the same time, if the viewer has a simple goal that can be predicted by the score report designer, then it might be valuable to create the appropriate graphical format.

When viewing line graphs, individuals primarily focus on *x*-*y* relationships (Carswell & Wickens, 1987; Shah & Carpenter, 1995; Shah & Freedman, 2009; Shah, Mayer, & Hegarty, 1999; Zacks & Tversky, 1999). In one particularly compelling finding, Zacks and Tversky (1999) found that viewers described trend information when viewing line graphs even when data were categorical. For example, when viewing a graph of heights of boys and girls, they might say, "as people become more male…", even though this is clearly not the correct interpretation. If multiple lines are depicted in the same graph, individuals will typically focus on comparisons between relative slopes of those lines (i.e., one line is increasing, another is decreasing), and pay less attention to the relative positions of those lines (Shah et al., 1999). In contrast, when individuals view bar graphs, they tend to compare the relative difference between bars that are grouped together (Shah & Freedman, 2009). Bar graphs are somewhat more neutral than line graphs, however, in that individuals are less likely to ignore differences between sets of bars than they are differences in relative position of lines.

Because line and bar graphs are so commonly used, research on comprehension of these displays has several potential implications for score reporting. Although bar graphs appear to be more common than line graphs for presenting score reports, there are cases where line graphs are heavily used. Some score reports plot subscore information in a line format, connecting subscore categories with lines. This choice is likely to lead to misinterpretation. Line graphs may be more commonly presented to teachers and administrators, and may be appropriate for presenting relative changes over time. However, such graphs may hide magnitude differences between groups or tests that are plotted as different lines. Bar graphs are another common format for presenting score information, but score report designers should be aware that viewers tend to focus on relative scores more than absolute scores because the relative differences in groups are very salient.

Pie charts are often used to present proportion data, and research has found that pie charts are often better for presenting relative proportions than divided bar charts (Spence & Lewandowsky, 1991). However, when absolute and magnitude information needs to be communicated, divided bar charts may be best (Kosslyn, 1994).

The effects of format suggest some basic guidelines for score reporting. If the goal of the graph designer is to make some information readily available to viewers (i.e., a students' strengths and weaknesses, the magnitude of the achievement gap, or the proportion of students who meet proficiency requirements), then they should utilize different formats depending on the nature of that information. In this chapter, however, we note that making some information salient relative to other information may have some important costs: Viewers might overinterpret or value some information relative to other information, they might not have the prior knowledge or graphical literacy skills to accurately interpret the data, and they may come away with an oversimplified or incorrect interpretation. In the next sections of the chapter, we discuss research about these possibilities and provide some suggestions to avoid such problems.

Salience of Graphs

Information presented in graphs is highly salient and persuasive—certainly more persuasive than the same information presented textually or numerically. Several studies support this idea. In a study demonstrating the power of graphics, for example, Fagerlin and her colleagues (Fagerlin, Wang, & Ubel, 2005) presented participants with anecdotes and numerical data regarding the likelihood that angina (chest pain) could be cured by balloon angioplasty. The statistical data was the same in all conditions (50% of individuals were cured). However, half the participants were given graphs depicting that data, and half were given the information in numeric form. Participants received four anecdotes about individuals who had undergone balloon angioplasty. In one condition, participants received four statistically representative anecdotes in which two patients were cured and two were not; in the other condition, participants were given four statistically nonrepresentative anecdotes in which only one of the four patients was cured. Typically, in making decisions about treatment options, people are highly influenced by anecdotes they hear. In other words, if they hear several anecdotes supporting one treatment option compared to another, they are more likely to pick that option. Fagerlin and colleagues found, however, that when given graphs representing statistical outcomes, participants were less likely to be influenced by anecdotes—that is, they made the same decisions regardless of the number of anecdotes supporting each treatment option. By contrast, when given the same statistical information in numerical form, individuals were more influenced by the anecdotes.

In another study demonstrating the power of graphics, participants were asked to make decisions regarding how much they would pay for products (better toothpaste or tires) that would reduce risks (of tooth decay or tire blowouts, respectively; Chua, Yates, & Shah, 2006). When the risk information was presented graphically, individuals stated that they would pay more to reduce risk than when risk information was presented numerically. Chua et al. (2006) found, further, that this decision was primarily caused by the fact that participants reported greater affective responses to the "risk" when the information was presented graphically. They also discussed other factors that may have played a role, including the idea that graphs cue the viewer that the data are scientific (e.g., Smith, Best, Stubbs, Johnston, & Archibald, 2000).

Although the discussion above suggests that graphs can have particular salience and persuasive power, numbers themselves may have salience relative to general qualitative statements such as "high ability," "proficient," and so forth. In a recent book, Charles Seife (2010) made exactly this point. He provided numerous examples in which providing an actual number, even an estimate, led individuals to overly rely on that number. One anecdote he shared was of a museum guide who, when asked how old an artifact was, stated that it was 65 million and 35 years old. When asked how the guide knew that number so precisely, he stated that when he first started working at the museum 35 years ago, a scientist told him the artifact was 65 million years old. Thus, he added 35 to the scientist's earlier estimate. This anecdote illustrates how a noisy measurement can be taken too seriously—a real concern with score reporting. Because scores are noisy measurements and different scores (especially subscores) vary in reliability, there is a risk that the numerical score values may be taken too seriously. The research on graphs suggests that if that same information is presented graphically, the risk of overinterpretation is even higher.

One direct implication of research regarding the relative importance of graphically presented information is that score report designers should be thoughtful when deciding which information to present graphically, which information to present numerically, and which information to present qualitatively or categorically. In general, more reliable scores should be presented graphically, whereas subscores with less reliability might be better presented either numerically or qualitatively.

Individual Differences: Beyond Quantitative Abilities

Much previous work on graph and data interpretation has focused on individual differences in quantitative skills, including knowledge about graphs and graph formats (e.g., Pinker, 1990). For example, Freedman and Shah (in press) demonstrated that high- and low-skilled individuals differed in their interpretation of graphs—specifically, low-skilled individuals were more likely to focus on surface-level attributes, such as visual features. One implication of these findings regarding the importance of quantitative skills in graph interpretation is that score reports should be designed with the quantitative skills of the audience in mind. When this is done, comprehension is much improved. Research in medical decision-making, for example, has shown that risk communications tailored on individual numeracy significantly improved understanding for a group of low numeracy Americans (Fagerlin, Ubel, Smith, & Zikmund-Fisher, 2007). Furthermore, Zikmund-Fisher, Fagerlin, and Ubel (2008) found that breast cancer patients, regardless of numeracy levels, understood information presented in a two-option pictograph better than in a four-option bar graph, reportedly because the former graphical display required less cognitive effort to interpret. Similar results may be found in terms of score reports. For lower numeracy audiences, score reports should include less information overall and focus on information that is readily retrievable.

The comprehension of graphs is not just affected by numeracy skills; rather, the comprehension of graphs (and other visual displays) is substantially knowledge-driven (see Figure 2; Kriz & Hegarty, 2007). Familiarity with the content of the information being depicted can have a large influence on comprehension (Canham & Hegarty, 2010; Shah & Freedman, 2009). We found, for example, that when data depicted information familiar to viewers, they were better able to draw appropriate inferences from graphs. In contrast, when data were unfamiliar, people primarily focused on salient visual information (Shah & Freedman, 2009).

One implication of such findings is that even when individuals report relatively high quantitative and graph comprehension literacy, they can have difficulty interpreting certain kinds of quantitative data. One study found, for example, that parents who self-reported having good graph reading skills nonetheless had difficulty interpreting the relationship between a child’s height and weight and what the percentiles represented (Ben-Joseph, Dowshen, & Izenberg, 2009). Furthermore, parents who examined a normal growth curve for short children mistakenly thought that the short children probably had major health problems.

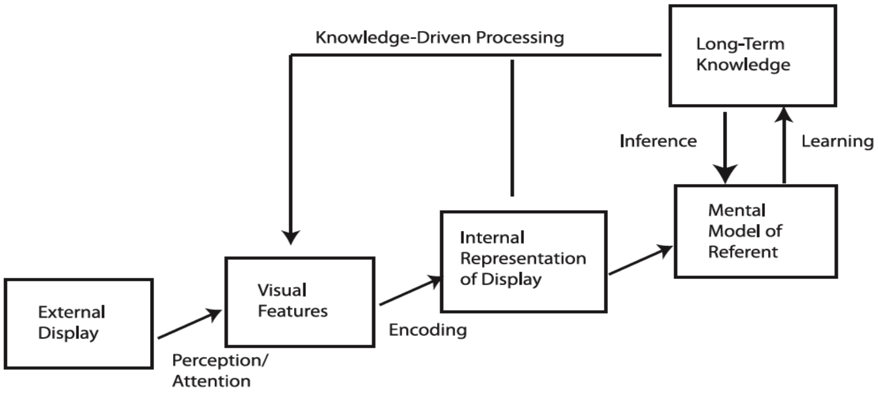


***Figure 2.* Model of visual display comprehension. From "Top-Down and Bottom-Up Influences on Learning From Animations," by S. Kriz and M. Hegarty, 2007, *International Journal of Human-Computer Studies*, *65*, pp. 911–930. Copyright 2007 by Elsevier. Reprinted with permission.**

Domain knowledge may similarly affect graphs in score reports. Consider a parent who understands how to read bar graphs and has high statistical literacy, but does not understand what subscores on the Wechsler Intelligence Scale for Children (WISC) test mean. Such a parent, when viewing Figure 3, will immediately note the anomalous short bar and become concerned about his or her child's performance. Indeed, a highly educated, statistically literate parent who saw similar scores for his child contacted one of us with concern about his child's "coding" ability. Understandably, he viewed the test as a measure of strengths and weaknesses of his child and sought to address the weaknesses. A parent with more knowledge about the test and the ability to interpret the overall score would be much less likely to be concerned about his child's visual-motor coordination skills, in light of her matrix reasoning scores. A clear implication of this example and, the role of domain knowledge in general, is that displays must use terms and quantitative variables familiar to parents. While the typical solution in score reports is to provide some text to explain different subscores, additional information would be useful, such as corresponding examples for each of the subscores and information regarding their predictive validity (or lack thereof).

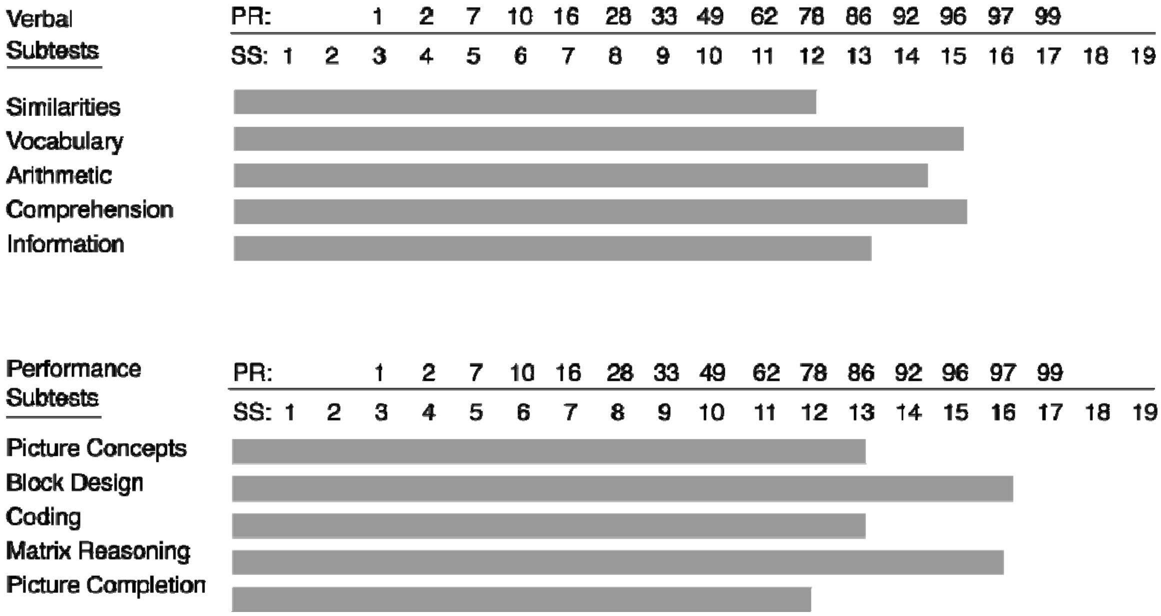


Figure 3. **Sample score report received by a parent in 2009, depicting performance on an intelligence test.**

Although yet to be examined in the context of evaluating and interpreting score reports, individual thinking dispositions have been shown to affect how people respond to information. Work examining the need for cognition, for example, finds that individuals vary in the level of pleasure they get from effortful analytic activity. For instance, individuals with a high need for cognition gravitate towards tasks requiring mental effort whereas those with a low need for cognition tend to favor simple tasks to complex ones. In one study, participants were asked to complete a mundane task but were given simple or complex instructions. Participants with a high need for cognition preferred the complex instructions over the simple instructions, where the opposite was the case for those with a low need for cognition (Cacioppo & Petty, 1982). Work by Nussbaum (2005) also found that a high need for cognition was related to producing more arguments in a persuasion task. Additionally, individuals with a high need for cognition are better able to discriminate between strong and weak arguments when evaluating persuasive texts, compared to those with a low need for cognition, and they also have better memory for arguments presented in the text (Cacioppo, Petty, & Morris, 1983). Research suggests that one mechanism for need for cognition may be motivation. See, Petty, & Evans (2009) gave participants equivalent sets of messages to evaluate, but half were told that the messages contained “technical wording” whereas the other half were told they contained “elementary wording.” For the participants in the technical wording manipulation, need for cognition was positively associated with reporting higher motivation towards the task.

The implication for score reporting is that individuals who have high need for cognition and engage in critical thinking may recognize that a test is a one-time sampling and that small differences on subscales are not very meaningful. In contrast, someone who focuses on the surface-level visual display of the same report may just focus on a single subscale and be concerned (or overly pleased) about performance, even though the differences on various subscales were small.

The nature and relevance of tasks can also influence the role need for cognition plays in how individuals respond to information. Our own work suggests that when individuals are instructed to critically evaluate data, they are more likely to do so and individual differences in need for cognition are not associated with providing more critical evaluations (Rodriguez, Shah, & Ng, 2010). With respect to score reporting, it may be beneficial to prime the viewer of score reports to critically evaluate the information provided. Our study suggests that even a reminder to do so is helpful.

Increasing Engagement With Score Reports

Feeling personally involved with a message stimulates attention and interest. When individuals think about information that has significant personal relevance, those with low need for cognition engage the same amount of argumentative responding as those with a high need for cognition (Axsom, Yates, & Chaiken, 1987). One recent approach to increasing personal interest is to tailor communications to satisfy each individual’s personal goals and encourage deeper processing of the message, thereby improving comprehension and memory.

Tailoring components of graphical displays to facilitate understanding for a variety of audiences strengthens the encoding step for each individual (Kriz & Hegarty, 2007), resulting in better internal representations of the display and, ultimately, greater long-term knowledge. In some cases, this can even have an effect on behavior change. A functional magnetic resonance imaging (fMRI) study using smokers showed that high-tailored messages resulted in increased activity in the medial prefrontal cortex, an area of the brain thought to be responsible for processing related to the self, when compared with low-tailored messages (Chua, Liberzon, Welsh, & Stecher, 2009). Activity in response to high-tailored messages was predictive of which smokers would quit, supporting the idea that self-relevance is an important component for messages encouraging behavior change. Other evidence has shown that self-related processing consistently results in superior memory across studies compared to other encoding strategies, such as semantic and other-referent (Symons & Johnson, 1997). A meta-analysis revealed that the greater the number of theoretical concepts tailored on, the greater the impact of tailoring (Noar, Benac, & Harris, 2007).

Tailoring score reports could be done in a number of ways. Possible dimensions include literacy, numeracy, education level, need for cognition, goal sets, perceived importance, and self-efficacy. Score reports are seen by a number of different people, and each recipient may be viewing it with a very different goal. Parents may be most concerned with how their child is performing in comparison to other children their age and may want to see a breakdown of subscores to understand how the overall score was computed. Students may just want to know that they are doing average, above average, or below average. Teachers and principals may be most concerned with the specific components of a standardized test that students are doing the best and worst on, to identify skills that are in need of extra attention. Individuals with advanced education and a higher need for cognition may want more detailed statistics, whereas those with low numeracy and low education may shrink from percentages and standard deviations, satisfied by simple graphical displays.

Desirable Difficulties in Graphs and Score Reporting

Displaying the important score information is not merely a matter of emphasizing a single fact over less important ones. Rather, research in graph comprehension as well as learning and judgment suggests that effectively communicating a concept or pattern is best characterized as a trade-off. Easing processing effort is important in certain situations, yet an effective display is often one that cognitively engages the viewer to process the information more deeply. The example at the beginning of the chapter highlights this issue. Because Figure 1 is relatively easy to interpret (i.e., the information that the achievement gap is reduced is highly salient), viewers might not realize that this “easy” interpretation is actually questionable.

In general, however, standard guidelines for graphic design focus on the importance of reducing cognitive effort. Specifically, current recommendations include reducing visual search times and offloading inference tasks to visual perception rather than logical thought (Larkin & Simon, 1987), reducing the sequence of eye fixations needed to encode a specific bit of information (Casner & Larkin, 1989), and avoiding redundancy or using the same modality to represent various types of information in the same display (Chandler & Sweller, 1991). Other work demonstrates how these goals might be accomplished by using perceptual groupings to highlight relevant trend information (Shah, Mayer, & Hegarty 1999), increasing the ratio of data-to-ink (Gillan & Richman, 1994; Neisser, 1963; Olzak & Thomas, 1986), or using graph formats that rely on visual judgment types demonstrated to be more effective for viewing quantitative information, such as the position-length judgments supported by bar graphs (Cleveland, 1985; Cleveland & McGill, 1984).

Despite the potential benefits of cognitive efficiency, empirical evidence suggests that information that is more difficult to process is actually better understood and remembered. For example, it is widely considered inappropriate to make 3-D bar graphs when depicting two-variable data because the perceptual processes are more difficult and error prone with 3-D graphs. At the same time, however, 3-D graphs are frequently preferred by viewers and can lead to better memory of information (Levy, Zacks, Tversky, & Schiano, 1996). In a related example, the standard recommendation is that whenever possible, information should be labeled so that it is easy to identify and keep track of referents (Kosslyn, 1994). In a study we recently conducted (Shah, Freedman, & Miyake, 2011), we asked participants to describe and answer questions about line graphs that depicted complex, multivariate data—half of the time the graphs had labeled lines, and the other half of the time the graphs had legends indicating which line was associated with each variable. Participants were faster to answer questions when lines were labeled, supporting the cognitive efficiency argument, which postulates that labeled lines are easier to read than legends. In fact, this is the recommendation typically made by graphic design handbooks. We found that when viewers were answering questions from memory, however, they were better in the legend condition. Furthermore, they were also better at making inferences about main effects depicted in the graphs in the legend condition than in the label condition.

One explanation for our results and others is that individuals may actually benefit from “desirable difficulties” in information presentation (Bjork & Bjork, 2011). Difficult displays require deeper, more active processing of information, which, in turn, can yield better comprehension and memory. Other studies that support this idea are ones that find animations to yield worse memory and comprehension than static displays of the same information. Whereas static displays require active processing, such as mental animation, display animations are more likely to be viewed passively (Hegarty, 2004; Hegarty, Kriz, & Cate, 2003).

Fonts are an easy-to-manipulate perceptual variable with demonstrable effects on comprehension. Studies vary the clarity of the font in which a questionnaire is printed, from very clear fonts like Times New Roman or Arial to difficult-to-process fonts like Haettenschweiler or Impact, to show that, in many cases, harder-to-process fonts improve comprehension of target information (see Alter & Oppenheimer, 2008; Alter, Oppenheimer, Epley, & Eyre, 2007; Novemsky, Dhar, Schwarz, & Simonson, 2007; Reber & Zupanek, 2002; Simmons & Nelson, 2006a, 2006b). One theory explaining the improvement stems from the fact that erroneous, intuitive, or heuristic (System 1) reasoning processes are less likely to be corrected under certain conditions, such as when people respond quickly (e.g., Bless & Schwarz, 1999; Chaiken, 1980; Petty & Cacioppo, 1986), but are more likely to be corrected when people are held accountable for their decisions (Tetlock & Lerner, 1999) or when disfluent experiences are used to induce more careful, analytical (System 2) reasoning (Alter et al., 2007). More recent work by Oppenheimer and his colleagues (Diemand-Yauman, Oppenheimer, & Vaughan, 2010) extends the findings on effect of font to memory and recall, demonstrating that a disfluent font (Haettenschweiler, Monotype Corsiva, and Comic Sans Italicized) leads to higher scores on classroom assessment tests.

The implication of the “desirable difficulties” perspective is that, in some cases, a graph design that introduces obstructions to purely passive processing of visual information may be beneficial. Consider, for example, the graph in Figure 1 again. If the information was presented in numeric form, rather than an easy-to-interpret graph, with sample size as salient as the mean score of each group, the viewer would have to mentally compute average performance for the White and African American students over time. In the process of doing so, however, he or she would be forced to attend to the information about the sample size. The final suggestion for score reporting, then, is that data presented in less-processed formats may lead to more initial difficulty in comprehension, but also fewer misinterpretations.

Conclusion

Score reports present quantitative information to different audiences about the scores of individuals or groups (i.e., classroom, schools, districts). Yet for a variety of reasons, interpreters of score reports may not form a complete and accurate understanding of the information presented. Psychological research on graph and visual-display comprehension points to several reasons why this might be the case: the salience and perhaps overemphasis of information presented numerically or graphically; the viewer's statistical and graphical literacy skills, domain knowledge, and dispositions; and the extent to which individuals deeply process information rather than merely attend to superficial visual features. To avoid these problems, designers of score reports should present only reliable, overall information graphically (using texts or tables to present subscore information), provide different levels of information regarding the content of tests whose score reports are being presented (i.e., sample problems, definitions), and develop displays that support active engagement (e.g., tailored displays, displays that do not yield a simple visual process but require some inferences and thought).

Teaching Teachers About Test Score Interpretation: The ITEMS Project[1]

Rebecca Zwick

ETS, Princeton, New Jersey

Jeffrey C. Sklar

California State Polytechnic University, San Luis Obispo

Abstract

This gap in assessment literacy was the impetus for the Instructional Tools in Educational Measurement and Statistics for School Personnel (ITEMS) project, which was based at the University of California, Santa Barbara, between 2004 and 2008 and was funded by the National Science Foundation. During the course of the project, our research team developed three web-based videos intended to improve the assessment literacy of K-12 educators by teaching educational measurement and statistics concepts, as applied to test score interpretation. The instructional videos were not designed as a replacement for an entire course, but rather as a professional development activity for teachers and school administrators or as a coursework supplement for students in teacher education programs. The effectiveness of the modules was evaluated through the administration of quizzes and through an independent program evaluation. The project is described in detail by Zwick et al. (2008); pedagogical aspects are discussed by Sklar and Zwick (2009).[2]

The current report describes the design, implementation, and results of the project, with a focus on the instructional approaches incorporated in the video modules.

Key words: score interpretation, assessment literacy, teacher professional development, teacher education

Project Overview

In each of three successive school years, the ITEMS project team developed, evaluated, and publicized a single video module. In the fall, we created the module (20 to 25 minutes in duration), along with a short quiz (14–20 multiple-choice items) that was geared to the module's content. The module and quiz were modified based on pilot data and on input from our project advisory committee, which consisted of teachers and school administrators, as well as university experts in human-computer interaction, multimedia learning, cognitive psychology, teacher education, educational technology, theoretical statistics, and math and statistics education.

In the winter and spring, we collected data on the module's effectiveness. The module was not publicly available during this time period; it could be accessed only by those with a project-assigned password. Educators participated in the project by logging into the project website. Participants first completed a background survey and then were randomly assigned (via a computerized "coin flip") to one of two conditions: In one condition, the module was viewed before the quiz on the module's content was administered; in the other, the quiz was administered first. Participants received a $15 (electronic) gift card from Borders and, in the later portion of the project, had the option of printing out a personalized completion certificate.

In the summer, we analyzed the quiz data to evaluate module effectiveness. By comparing participants from the two conditions—those who answered the quiz after viewing the module and those who answered the quiz before viewing the module—we were able to test the hypothesis that those who viewed the module first were better able to answer the quiz questions. Results are discussed in a later section.

Two additional data collection efforts occurred subsequently. Participants willing to be followed up took the quiz a second time, one month after their initial participation, to provide a measure of retention. In addition, an independent evaluator used interviews and surveys to obtain the perspectives of school personnel regarding the utility and effectiveness of the materials and to solicit suggestions for improvement. These phases of the project are discussed in Zwick et al. (2008) and Sklar and Zwick (2009).

After all data had been collected, we made the module publicly available on our Website, along with supplementary materials, including a glossary, formulas, and examples. We also distributed free CDs or DVDs containing the materials to educators who requested them.

Principles of Module Development

Module 1, “What’s the Score?” was developed in 2005. It described test score distributions and their properties (mean, median, mode, range, standard deviation), types of test scores (raw scores, percentiles, scaled scores, and grade-equivalents), and norm-referenced and criterion-referenced score interpretation. Module 2, “What Test Scores Do and Don’t Tell Us” (2006) focused on the effect of measurement error on individual student test scores, the effect of sample size on the precision of average scores for groups of students, and the definition and effect of test bias. Module 3, “What’s the Difference?” (2007) discussed data aggregation issues and addressed the interpretation of test score trends and group differences.

The instructional modules used realistic test score reports as a basis for explaining concepts and terminology. In computer-based learning environments, it has been found that individuals who are presented with material via an animated pedagogical agent demonstrate better learning outcomes than those who are presented with the material via on-screen text and static graphs (Moreno, Mayer, Spires, & Lester, 2001). Therefore, the modules made liberal use of graphics, including computer animation. The modules used cartoon characters, representing teachers, students, a superintendent, parents, and reporters, to present and discuss concepts. This decision led to decidedly mixed comments from participants. Some stated that the cartoons added just the right light and whimsical touch to material that can sometimes be dry, while others found the approach to be distracting, or, in a few cases, condescending. Additional research is needed to identify the characteristics of audiences and learning contexts associated with the successful use of cartoon characters as pedagogical agents.

In designing the modules, we sought to incorporate established principles from the cognitive psychology literature, including the following:

- **Multimedia principle:** Concepts were presented using both words and pictures. Research has shown that “…human understanding occurs when learners are able to mentally integrate visual and verbal representations” (Mayer, 2001, p. 5).
- **Contiguity principle:** Auditory and visual materials on the same topic were, whenever possible, presented simultaneously, rather than successively, and words and corresponding pictures appeared on the screen together rather than separately. Materials that incorporate these principles of temporal and spatial contiguity have been shown to enhance learning (Mayer, 2001, pp. 81–112).

- **Prior knowledge principle:** The modules were designed to "use words and pictures that help users invoke and connect their prior knowledge" to the content of the materials (Narayanan & Hegarty, 2002, p. 310). For example, while participants may be unfamiliar with the term, "measurement error," most have had the experience of weighing something (possibly themselves) twice and getting disparate results. Analogies and metaphors have been shown to enhance mathematical learning (English, 1997).
- **Personalization principle:** An informal conversational style was used in the modules; this has been shown to enhance learning (Mayer & Moreno, 2002), perhaps because "learners may be more willing to accept that they are in a human-to-human conversation including all the conventions of trying hard to understand what the other person is saying" (Mayer, 2003, p. 135). In keeping with this principle, formulas are not used in the instructional modules. (They are included only in the supplementary materials posted on the web.) With regard to the presentation of technical material, our philosophy was much the same as that of the statistics textbook authors Freedman, Pisani, Purves, and Adhikari (1991, p. xiii), who stated, "Mathematical notation only seems to confuse things for most people, so we [explain statistics] with words, charts, and tables—and hardly any x's or y's … What [people] really need is a sympathetic friend who will explain the ideas and draw the pictures behind the equations. We are trying to be that friend…"

Pedagogical Challenges: Some Examples

In this section we discuss the instructional and pedagogical approaches that were used in the modules along with specific examples. We used both static graphs and dynamic images to illustrate mathematical procedures and statistical concepts, created realistic test score reports to illustrate measurement principles, and used analogies to help viewers connect their prior knowledge to new concepts. Conveying mathematical or statistical information without using formulas was by far our biggest pedagogical challenge. We sought to replace traditional mathematical formulas with dynamic images and graphics that could represent mathematical operations. Some examples of our pedagogical approaches follow.

In Module 1, a dynamic graphical sequence was used to introduce the idea of a distribution of test scores, a concept that was unfamiliar to many teachers, according to our preliminary research. We attempted to connect the abstract idea of a distribution to a more literal representation: A teacher was shown throwing test score reports into labeled bins corresponding to test score intervals. In the final image, the test score distribution was represented by a histogram formed by the stacks of reports. The mean, median, standard deviation, and skewness of the distribution were then discussed.

In Module 2, we illustrated a device for conceptualizing measurement error that is often used in educational measurement textbooks. The idea is that a child takes a test repeatedly. His brain is magically purged of his memory of the test between testing occasions. For various reasons, he gets different scores each time, as illustrated in Figure 1. The viewer is asked to imagine that that the pictured child, Edgar, takes a test several times, magically forgetting the content of the test between administrations. On the first occasion, he misreads a question to which he knows the answer, getting it wrong; on the second, he guesses correctly on a question to which he does not know the answer; and on the third, he is accidentally given extra time on the test. For these reasons, he gets slightly different scores on each imaginary test administration.

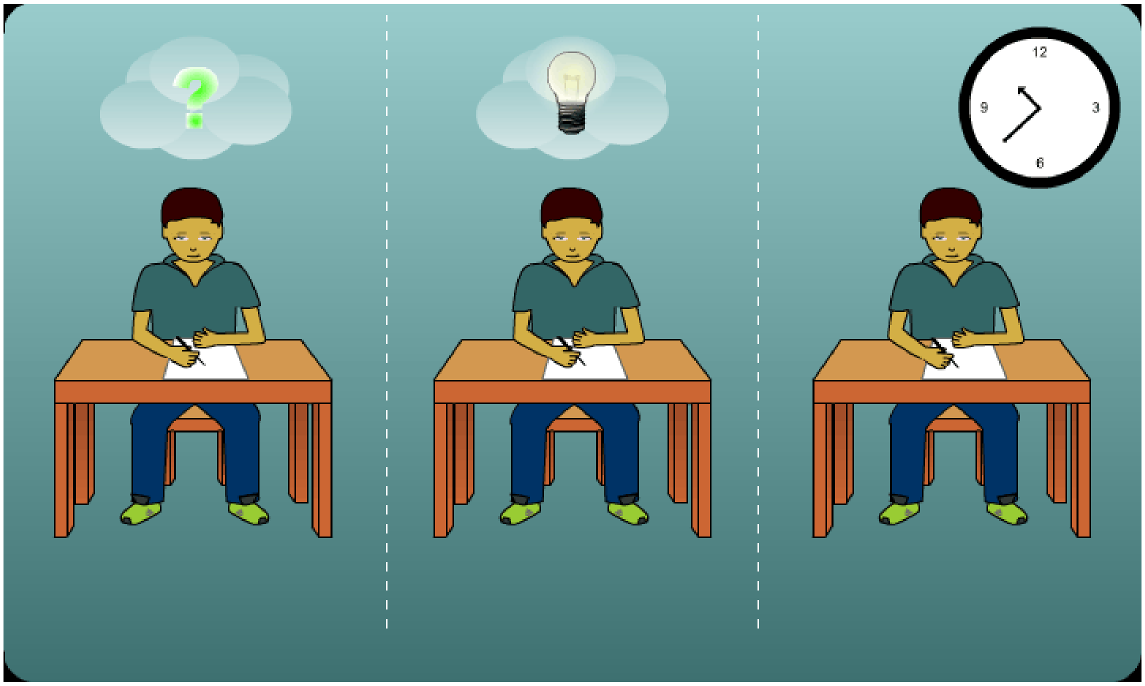
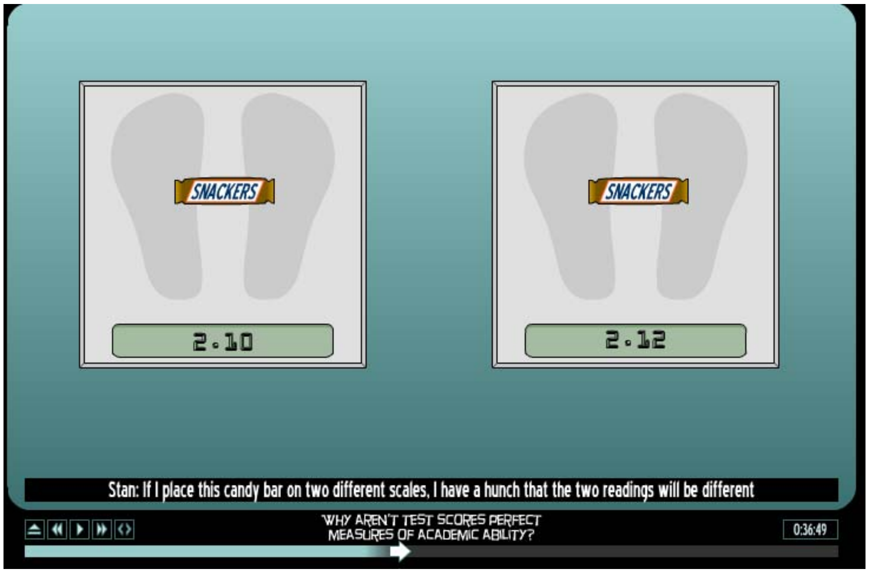


Figure 1. **Illustration of the effect of measurement error on a student's score.**

Note. Image by Graham Wakefield. From "Instructional Tools in Educational Measurement and Statistics (ITEMS) for School Personnel: Evaluation of Three Web-Based Training Modules," by R. Zwick et al., Educational Measurement: Issues and Practice, *27*, pp. 14–27. Copyright 2010 by the National Council on Measurement in Education. Used with permission.

Module 2 also included an illustrative analogy concerning measurement error outside of the realm of test scores, as shown in Figure 2. Two side-by-side scales are displayed, each weighing a candy bar, but showing two different weight readings. The scene illustrates that, because of imprecision in measuring capabilities, different results may be obtained on different measurement occasions. This phenomenon is similar to the imprecision involved in using educational tests to measure student skills.

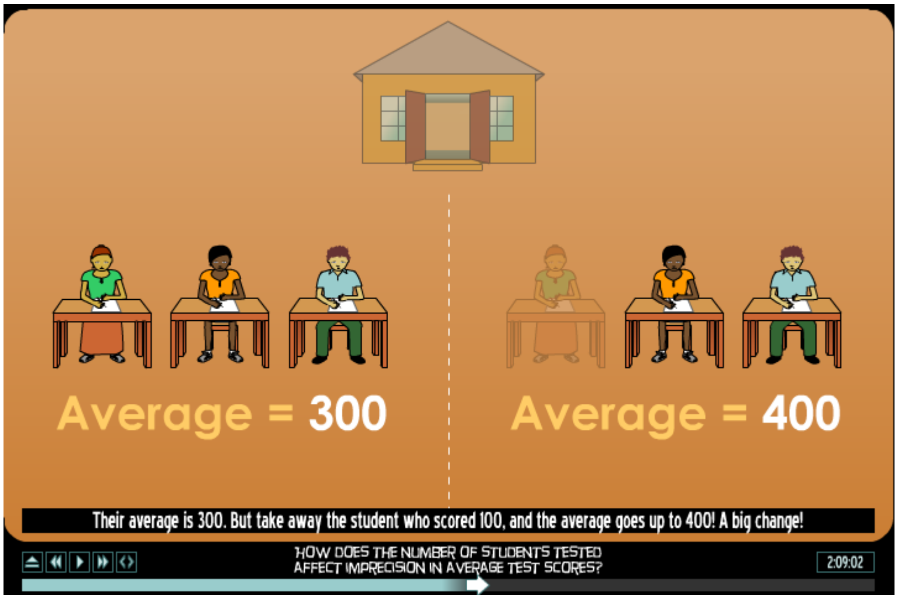

SNACKERS
2.10
SNACKERS
2.12
WHY AREN'T TEST SCORES PERFECT MEASURES OF ACADEMIC ABILITY?
0:36:49

Stan: If I place this candy bar on two different scales, I have a hunch that the two readings will be different

***Figure 2.* An illustration of the effect of measurement error on weight data.**

Note. Image by Graham Wakefield. Copyright 2010 by the Regents of the University of California.

Another topic addressed in Module 2 was sampling error. We wanted to convey in a simple, nontechnical way the idea that a mean based on a small sample is less trustworthy than one based on a large sample, other things being equal. We illustrated this by showing the effect of individual test scores on the average test score for a class or school. A group of students and their average test score was first displayed. The next image showed a particular student and his test score being removed. Then an image of the newly reduced group of students and their average test score was displayed. In the left panel of Figure 3, the average score for the three students was 300. In the right panel, we can observe the leftmost student, who had a very low score, fading from the image, and a new average test score appearing (Average = 400). From this scene, viewers can observe that if a class size is small, then one student with an extreme test score can have a large impact on the average score. Another sequence of images showed viewers that when the class size was large, removing a single student's score had little impact.



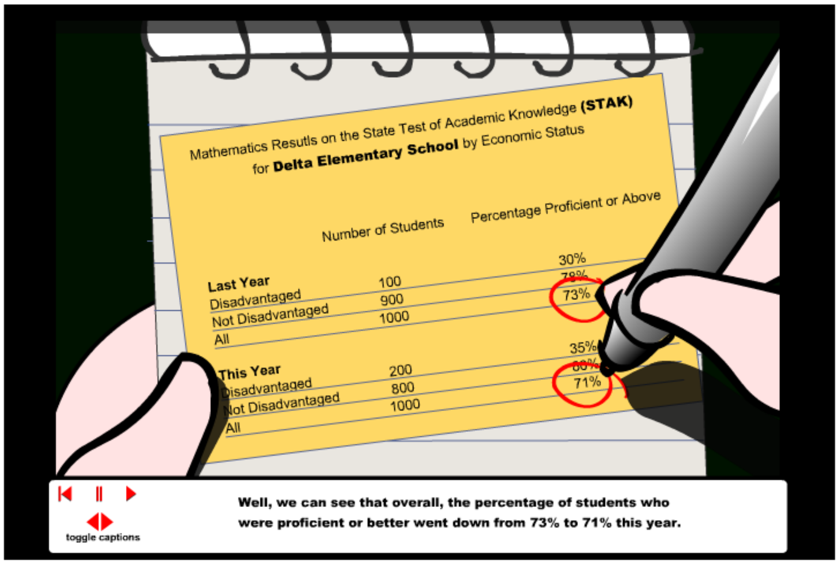
Their average is 300. But take away the student who scored 100, and the average goes up to 400! A big change!

HOW DOES THE NUMBER OF STUDENTS TESTED AFFECT IMPRECISION IN AVERAGE TEST SCORES?

***Figure 3.* Illustration of the effect of removing one student on the average score.**

Note. Image by Graham Wakefield. Copyright 2010 by the Regents of the University of California.

Another challenging statistical topic, known as Simpson's Paradox, was illustrated in Module 3. Simpson's Paradox, sometimes called the amalgamation paradox, occurs when the direction of an association between two variables is reversed when a third variable is controlled (see Utts & Heckard, 2004, for examples). Our goal was to illustrate this phenomenon with a specific and realistic example. In one Module 3 scene, the paradox was observed at a particular school, where the proficiency rate increased from 30% to 35% from one year to the next for students in an economically disadvantaged group and from 78% to 80% in the nondisadvantaged group. The overall proficiency rate for all students combined, however, decreased from 73% to 71% (see Figure 4). The reason for this apparent oddity is that the proportion of disadvantaged students increased from 10% to 20%, while the proportion of nondisadvantaged students decreased from 90% to 80%.



Mathematics Resutls on the State Test of Academic Knowledge **(STAK)** for **Delta Elementary School** by Economic Status

	Number of Students	Percentage Proficient or Above
Last Year		
Disadvantaged	100	30%
Not Disadvantaged	900	78%
All	1000	73%
This Year		
Disadvantaged	200	35%
Not Disadvantaged	800	80%
All	1000	71%

Well, we can see that overall, the percentage of students who were proficient or better went down from 73% to 71% this year.

Figure 4. **An illustration of Simpson's paradox in the context of test results.**

Note. Image by Cris Hamilton. Copyright 2010 by the Regents of the University of California.

Module Effectiveness

Average scores of the quiz-first and module-first groups were compared to determine the effectiveness of the modules. Data were analyzed for teacher education students, for school personnel, and for the combined group. Across all three modules, the teacher education students had an average age of 26 and an average of two years of teaching experience. The school personnel had an average age of 48 and an average of 17 years of experience. Overall, the majority of research participants were women.

Table 1 (on page 27) displays the means and standard deviations of the quiz scores for the quiz-first and module-first groups for all three quizzes, along with the sample sizes. In general, school personnel outperformed teacher education students, as measured by their average quiz score. However, the differences in average scores between the module-first and quiz-first groups, which provide an estimate of module effectiveness, were larger among teacher education students than among school personnel. The results of one-sided t-tests comparing the module-first and quiz first groups are shown in Table 1 for teacher education students, school personnel, and the combined group. Results for teacher education students and school personnel are also illustrated in Figures 5–7, pp. 28–30. Among teacher education students, the effect sizes due to Modules 1, 2, and 3 were .35, .84, and .24 standard deviation units, respectively, while the corresponding effect sizes were .28, .10, and .20 among school personnel. These results suggest that teacher education students benefited more than school personnel from the module presentations, particularly in Module 2 (see Figure 6, p. 29). Further details are provided in Zwick et al. (2008).

Conclusions

The goal of the ITEMS Project was to create short web-based presentations that would assist pre-service and in-service teachers, as well as school administrators with interpreting standardized test results. An evaluation of the effectiveness of these video modules showed that they had a positive impact, particularly in the case of teacher education students. The project received generally positive feedback from participants. For example, one educator called the materials "[v]ery helpful and right to the point. If I were a building principal … all of the staff would go through this until everyone really understood it." The modules were adopted for ongoing use in some districts and at least one teacher education program.

There were several challenges associated with developing the presentations. Not only was the material complex, but time was limited. Based on the lessons learned from the ITEMS project, Sklar and Zwick (2009) developed recommendations for designing Web-based instructional materials in educational measurement and statistics, including the following:

- Presentations should implement multimedia design principles.
- Topics should be presented in clearly partitioned scenes rather than one single continuous presentation.
- Complex mathematical equations and computations should be avoided
- Analogies should be used to invoke prior knowledge.
- Realistic mock-ups of test score reports should be used as illustrations.

Future research should focus on empirical investigations of these design features and instructional approaches. Research of this kind could serve to improve the quality of professional development tools in educational measurement and statistics, an important short-term goal. In the longer term, improvement of teacher qualifications in this area are unlikely to occur without changes in teacher licensing requirements in the area of assessment literacy, which, in turn, would spur the much-needed modifications in teacher education curricula.

Table 1

Means, Standard Deviations, and Sample Sizes, and t-Test Results for Quiz Scores

	Teacher education students					School personnel					All participants combined				
	Module-First		Quiz-First		*t*-test *p*-value	Module-First		Quiz-First		*t*-test *p*-value	Module-First		Quiz-First		*t*-test *p*-value
Module	Mean (SD)	Sample size	Mean (SD)	Sample size		Mean (SD)	Sample size	Mean (SD)	Sample size	*t*-test *p*-value	Mean (SD)	Sample size	Mean (SD)	Sample size	*t*-test *p*-value
1	13.1 (4.0)	33	11.7 (3.5)	35	.059	13.4 (3.2)	19	12.5 (3.2)	26	.198	13.2 (3.7)	52	12.0 (3.4)	61	.042
2	12.6 (3.2)	40	9.5 (3.7)	41	.000	12.7 (1.9)	11	12.5 (1.4)	12	.375	12.6 (3.0)	51	10.2 (3.5)	53	.000
3	6.5 (4.1)	8	5.5 (2.1)	6	—	11.2 (3.0)	10	10.4 (4.0)	9	—	9.1 (4.2)	18	8.5 (4.1)	15	.66

Note. The t-test p-values are the one-sided p-values corresponding to the t-test comparing the module-first and quiz-first groups. For Module 3, t-tests were computed only for the combined group of participants because of small sample sizes.

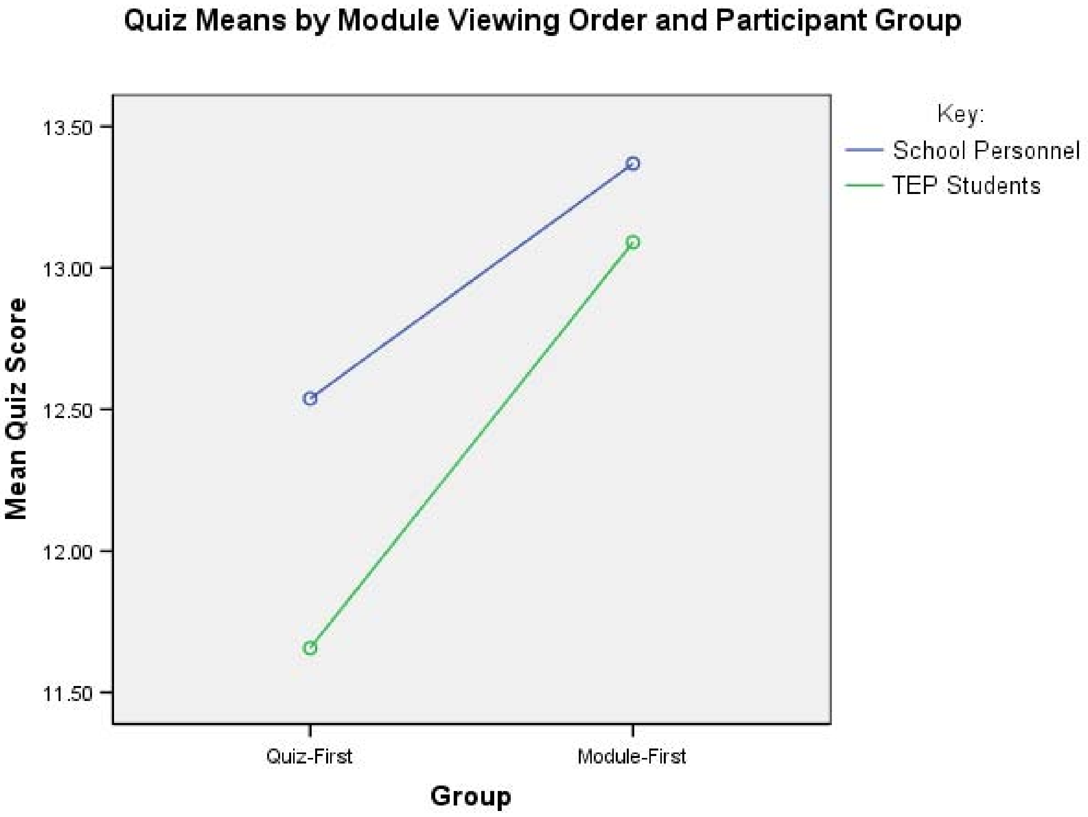


Figure 5. **Results for Module 1.**

Note. The number of items in the quiz was 20. See Table 1 for means, standard deviations, sample sizes, and *t*-test results. TEP = Teacher Education Program. From "Instructional Tools in Educational Measurement and Statistics (ITEMS) for School Personnel: Evaluation of Three Web-Based Training Modules," by R. Zwick et al., *Educational Measurement: Issues and Practice*, *27*, pp. 14–27. Copyright 2010 by the National Council on Measurement in Education. Used with permission.

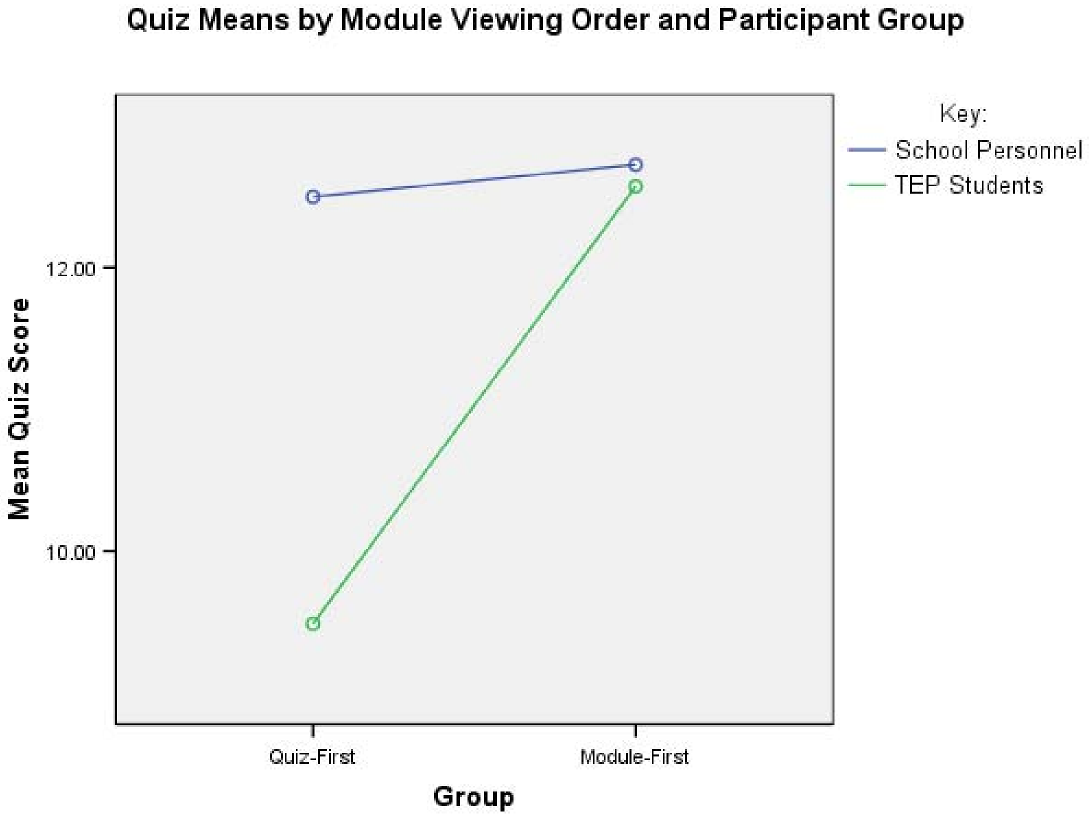


Figure 6. **Results for Module 2.**

Note. The number of items in the quiz was 16. See Table 1 for means, standard deviations, sample sizes, and *t*-test results. TEP = Teacher Education Program. From “Instructional Tools in Educational Measurement and Statistics (ITEMS) for School Personnel: Evaluation of Three Web-Based Training Modules,” by R. Zwick et al., *Educational Measurement: Issues and Practice, 27*, pp. 14–27. Copyright 2010 by the National Council on Measurement in Education. Used with permission.

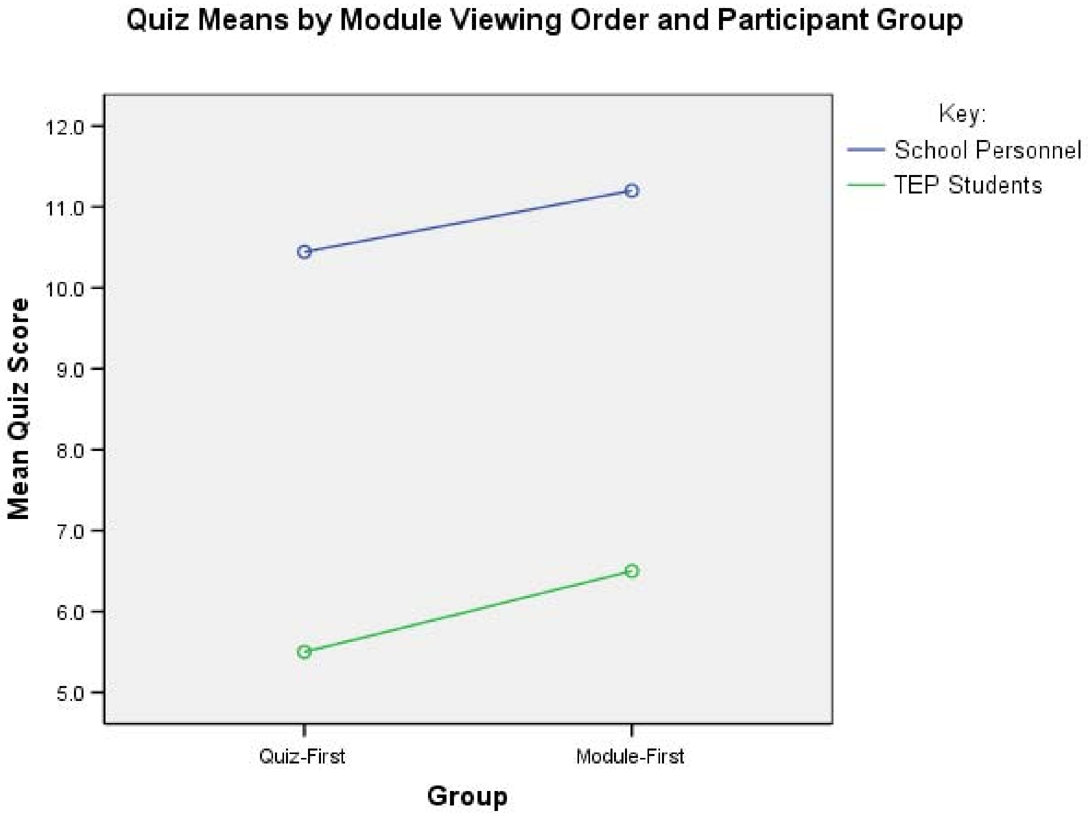


***Figure 7.* Results for Module 3.**

Note. The number of items in the quiz was 14. See Table 1 for means, standard deviations, sample sizes, and *t*-test results. TEP = Teacher Education Program. From "Instructional Tools in Educational Measurement and Statistics (ITEMS) for School Personnel: Evaluation of Three Web-Based Training Modules," by R. Zwick et al., *Educational Measurement: Issues and Practice*, *27*, pp. 14–27. Copyright 2010 by the National Council on Measurement in Education. Used with permission.

Notes

[1] This work was conducted when the first author was at University of California, Santa Barbara. This project was funded by the National Science Foundation (#0352519). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

[2] The two articles and all ITEMS videos, quizzes, and supplementary materials are available at http://items.education.ucsb.edu

Designing and Evaluating Score Reports for Particular Audiences

Diego Zapata-Rivera
ETS, Princeton, New Jersey

Abstract

Although principles for designing high-quality score reports have been proposed and professional standards indicate that test takers need to be informed about assessment results as well as the purpose of the assessment and its recommended uses, many of the score reports available do not effectively convey this score information for particular audiences. Our work seeks to design and evaluate score reports that clearly communicate useful assessment information to various educational stakeholders. This paper presents a framework for designing and evaluating score reports and describes our work on score reporting for three different audiences: teachers, administrators, and students.

Key words: score reporting, particular audiences, policymakers, administrators, teachers, students

Acknowledgments

I would like to acknowledge the members of the CBAL reporting team: Meg Vezzu, Waverely VanWinkle, Rebecca Zwick, Janet Stumper, Debbie Pisacreta, and Tom Florek for their excellent work on the project. I would like to extend our gratitude to the teachers, administrators, and students who participated in our studies. I also want to thank the members of the CBAL project for providing useful feedback and supporting this work. Finally, I would like to thank Eric Hansen, Christine Lyon, and Daniel Eignor for their comments and suggestions.

Background

Existing research on score reports indicates that teachers, school administrators, and policy makers 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, the purpose of the assessment and its recommended uses (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999); many currently available score reports do not effectively convey this information to particular audiences.

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

This paper describes a framework for designing and evaluating score reports that includes the following activities: (a) gathering assessment information needs from stakeholders, (b) reconciling these needs with the available assessment information, (c) designing various score report prototypes, and (d) evaluating these score report prototypes internally and externally. It also presents score reports for teachers, local and state-level administrators, and students that have been designed and evaluated following this framework.

Related Research

Relevant literature includes work on heuristics for creating reports that communicate the intended message to a particular audience (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). There is also evidence suggesting that teachers require and would benefit from additional training on basic educational measurement concepts required to understand information that is usually included in the score reports (Bennett & Shepherd 1982; Lukin, Bandalos, Eckhout, & Mickelson, 2004; Zapata-Rivera, VanWinkle, & Zwick, 2010; Zwick et al., 2008).

Deng and Yoo (2009) present an extensive list of score reporting resources that includes papers, guidelines, and sample score reports. Roberts and Gierl (2010) review current score reporting practices, propose a framework for developing score reports for cognitive diagnostic assessments, and showcase a score report for students in the domain of mathematics. The authors highlight 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) presents a series of principles for developing effective multimedia applications based on empirical evidence. These principles include: (a) *Various representations*: students learn better when both words and pictures are presented than when only words are used; when both words and pictures are used, students have a chance to create verbal and pictorial mental models and to construct links between them; (b) *Spatial contiguity*: students learn better when related words and pictures are placed near each other on the page or screen; this way learners do not need to dedicate cognitive resources to visually scan the page or screen, and the likelihood of keeping the words and pictures in working memory will increase; (c) *Coherence*: students learn better when extra, non-relevant material is excluded; extra material causes competition for limited cognitive resources and can be distracting; (d) *Redundancy*: students learn better when only given animation and narration than when given animation, narration, and text that is the same as the narration; pictures and written words share the same visual channel which can cause cognitive overload; (e) *Individual differences*: effects of multimedia design are stronger for learners with low levels of knowledge than for learners with high levels of knowledge; students with high levels of knowledge can rely on prior knowledge to make up for a lack of guidance in poorly designed presentations, while students with low levels of knowledge cannot. In addition, high spatial-ability students have the ability to combine visual and verbal content from a multimedia presentation, while low spatial-ability learners may not have the same ability.

We have designed score reports for particular educational stakeholders. These reports 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 (formative hypotheses are tentative statements about student performance that teachers can use in conjunction with other available evidence to inform instruction), performance levels, and scaled scores based upon availability of supporting evidence (e.g., see Appendix A). Reports for school administrators are tailored to respond to particular questions of interest to this audience. These reports usually include performance information aggregated at the grade, school or school district levels, information about particular subgroups, and comparisons with similar schools or school districts. Student reports may include (a) task-level information, (b) areas that show good performance or may need improvement, (c) performance levels, and (d) scaled scores.

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. It includes the following steps: (a) gathering assessment information needs, (b) reconciling these needs with the available assessment information, (c) designing various score report prototypes, and (d) evaluating these report prototypes internally and externally. Figure 1 depicts this framework graphically.

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 the form of a document called the prospective score report (PSR) that is used to gather client assessment 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).

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 sub-skills, possible performance levels, comparison, progress, and task-level information. The PSR provides a way for us to communicate our understanding of the reporting needs to content and measurement experts for their evaluation.

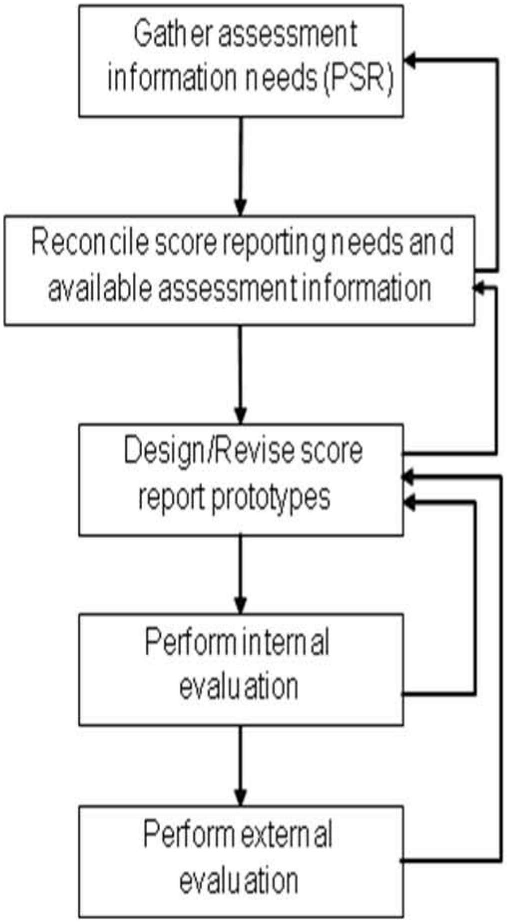

Gather assessment information needs (PSR)
Reconcile score reporting needs and available assessment information
Design/Revise score report prototypes
Perform internal evaluation
Perform external evaluation

***Figure 1.* A framework for designing and evaluating score reports 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 and how this information is presented in order to ensure that the intended audience receives the intended message while steering them clear of inappropriate uses.

Once an initial consensus has been achieved, various score reports can be designed following best score report design practices. It is worth mentioning that 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 information” and “Design/revise score report prototypes” in Figure 1).

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 external representations should be followed. For example, work by Fast (2002), Goodman and Hambleton (2004), Underwood, Reshetar, and Leahy (2006), and Hattie (2009).

The use of pre-existing score report templates that have previously been evaluated can facilitate this process. However, new elements need to be designed to incorporate report components that were not initially covered. Several score report variants are created to explore alternate representations. These variants may include different graphical representations, layouts, interpretive text, interactive components, etc.

Data to populate the score reports may be created to resemble actual data or 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 internally first with the help of content, measurement, usability, and accessibility experts. Information gathered from experts is used to refine, create, or abandon score report variants. Resulting score reports are evaluated externally by conducting qualitative and quantitative studies with the intended audience. Data acquired 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.

The next sections describe prototype score reporting systems created for teachers, administrators, and students.

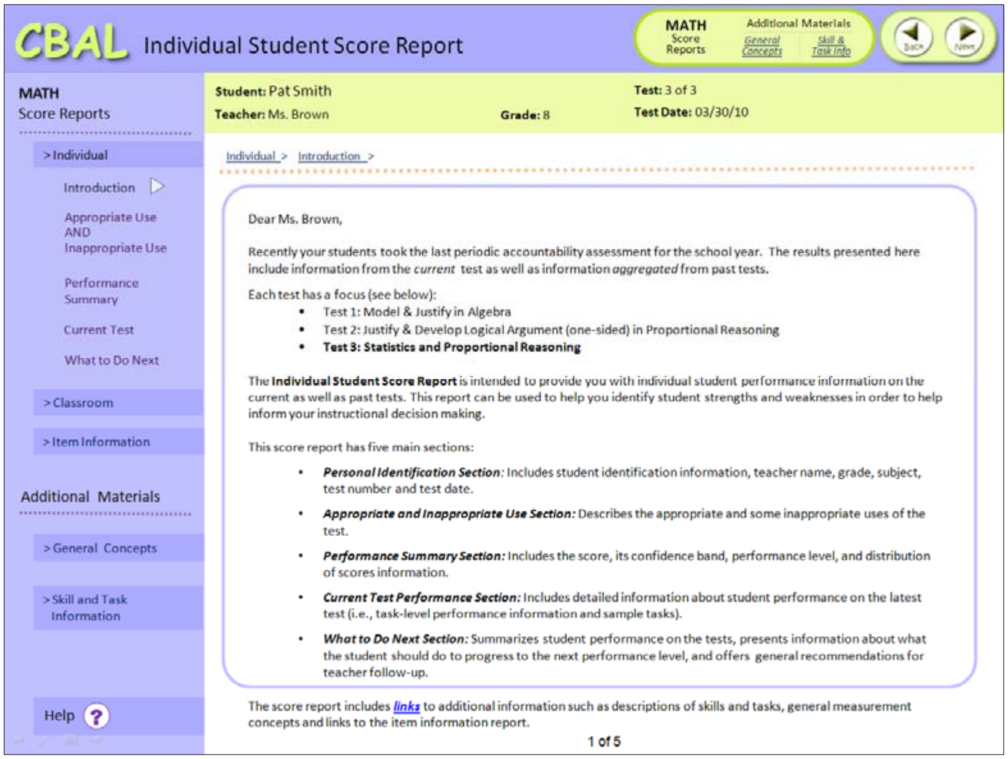
Score Reports for Teachers

Three types of score reports for teachers have been developed for CBAL: 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). The report 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, skill, and task information are available though the vertical navigation pane as well as through the underlined hyperlinks integrated into the score report.

In addition to the individual student score report, two other types of score reports are available for teachers: a classroom score report and an item information report. The classroom 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 by clicking on an individual’s name.

The item information report 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 as links.



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.

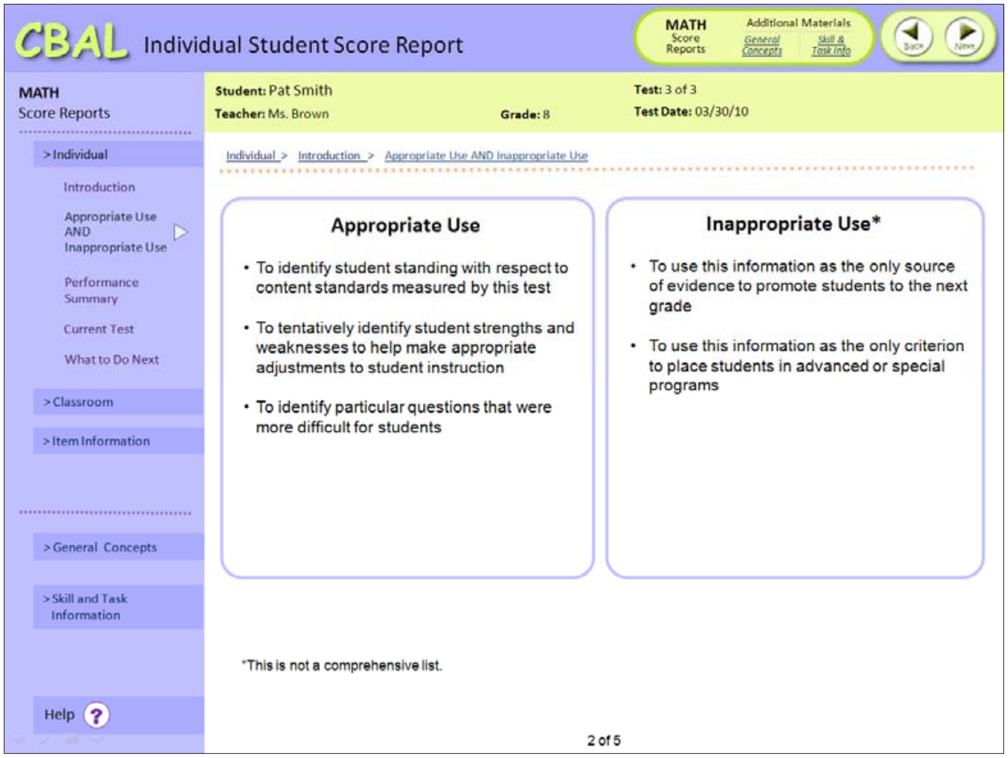
This score report has five main sections:

- ***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 (p. 1 of 5).**



Student: Pat Smith

Teacher: Ms. Brown **Grade:** 8

Test: 3 of 3

Test Date: 03/30/10

Individual > Introduction > Appropriate Use AND Inappropriate Use

Appropriate Use

- To identify student standing with respect to content standards measured by this test
- To tentatively identify student strengths and weaknesses to help make appropriate adjustments to student instruction
- To identify particular questions that were more difficult for students

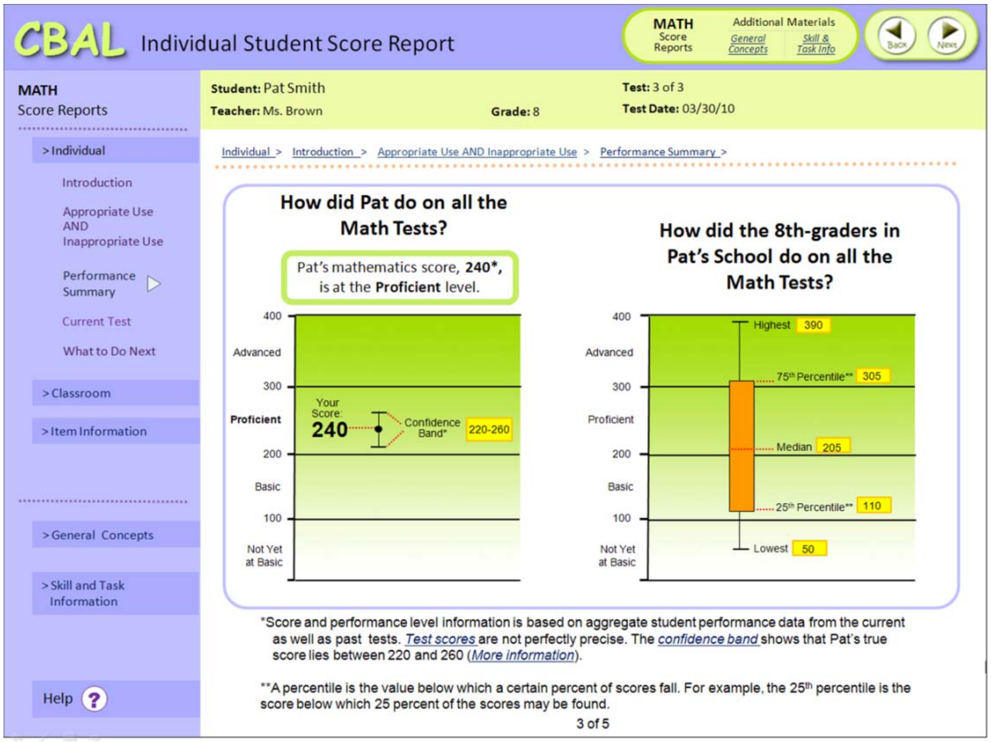
Inappropriate Use*

- To use this information as the only source of evidence to promote students to the next grade
- To use this information as the only criterion to place students in advanced or special programs

*This is not a comprehensive list.

2 of 5

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



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

How did Pat do on all the Math Tests?

Pat's mathematics score, **240***, is at the **Proficient** level.

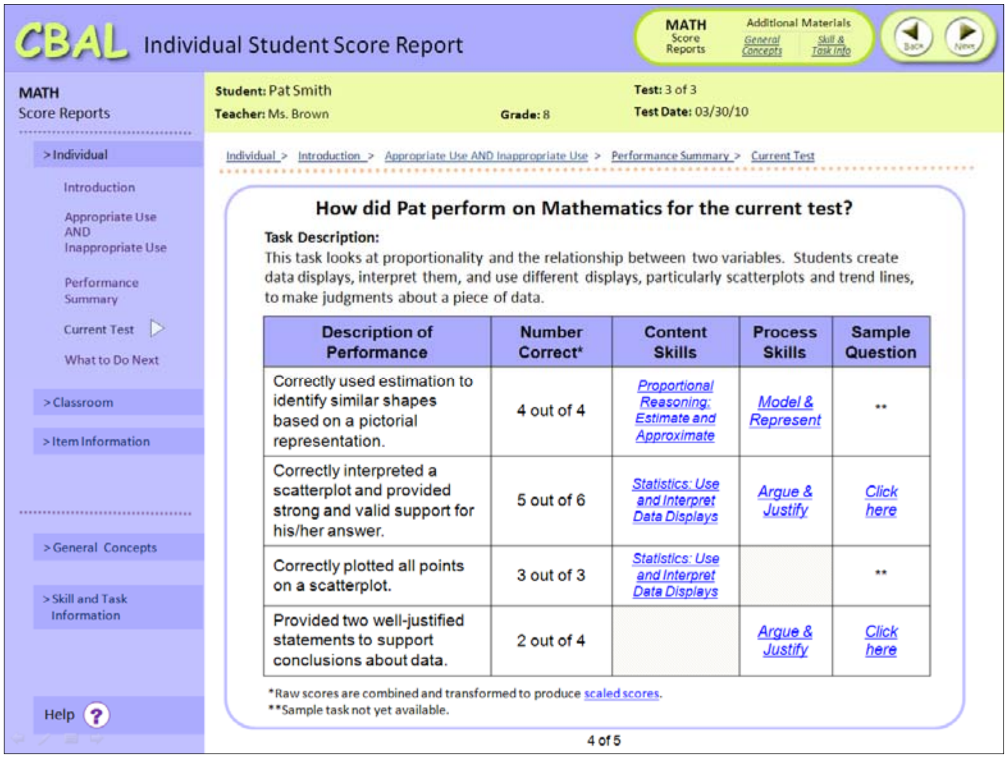
How did the 8th-graders in Pat's School do on all the Math Tests?

*Score and performance level information is based on aggregate student performance data from the current as well as past tests. Test scores are not perfectly precise. The confidence band shows that Pat's true score lies between 220 and 260 (More information).

**A percentile is the value below which a certain percent of scores fall. For example, the 25th percentile is the score below which 25 percent of the scores may be found.

3 of 5

Figure 4. **Performance summary (p. 3 of 5).**



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

How did Pat perform on Mathematics for the current test?

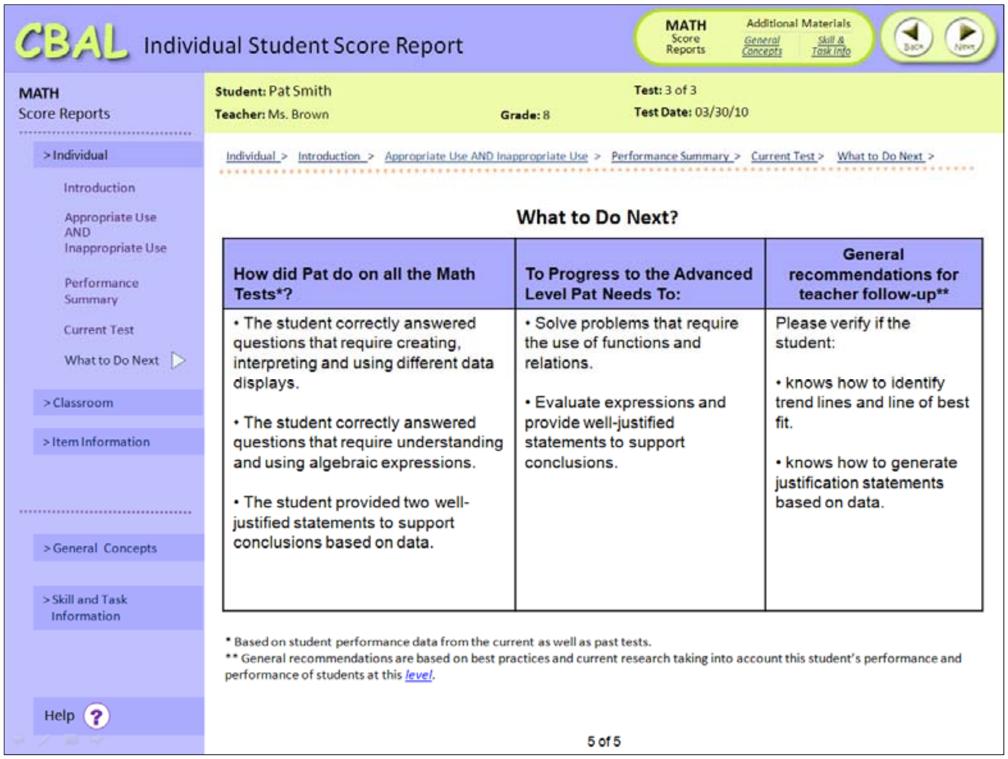
Task Description:
This task looks at proportionality and the relationship between two variables. Students create data displays, interpret them, and use different displays, particularly scatterplots and trend lines, to make judgments about a piece of data.

Description of Performance	Number Correct*	Content Skills	Process Skills	Sample Question
Correctly used estimation to identify similar shapes based on a pictorial representation.	4 out of 4	Proportional Reasoning: Estimate and Approximate	Model & Represent	**
Correctly interpreted a scatterplot and provided strong and valid support for his/her answer.	5 out of 6	Statistics: Use and Interpret Data Displays	Argue & Justify	Click here
Correctly plotted all points on a scatterplot.	3 out of 3	Statistics: Use and Interpret Data Displays		**
Provided two well-justified statements to support conclusions about data.	2 out of 4		Argue & Justify	Click here

*Raw scores are combined and transformed to produce scaled scores.
**Sample task not yet available.

4 of 5

Figure 5. **Current test performance (p. 4 of 5).**



Individual Student Score Report

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

Individual > Introduction > Appropriate Use AND Inappropriate Use > Performance Summary > Current Test > What to Do Next >

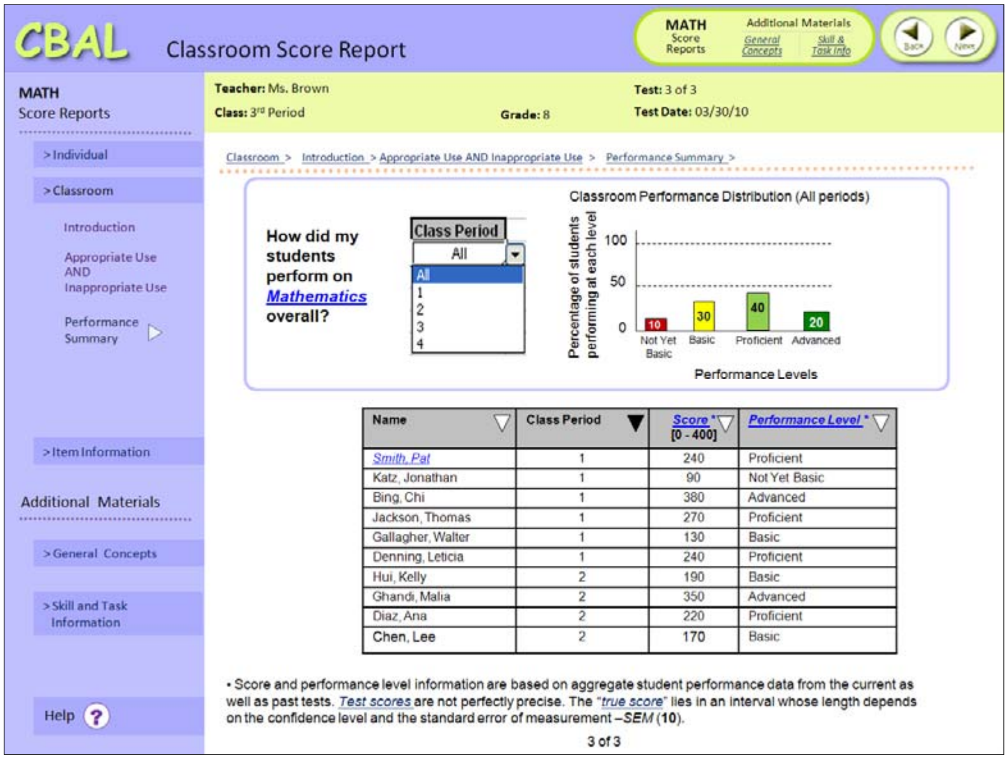
What to Do Next?

How did Pat do on all the Math Tests*?	To Progress to the Advanced Level Pat Needs To:	General recommendations for teacher follow-up**
• The student correctly answered questions that require creating, interpreting and using different data displays. • The student correctly answered questions that require understanding and using algebraic expressions. • The student provided two well-justified statements to support conclusions based on data.	• Solve problems that require the use of functions and relations. • Evaluate expressions and provide well-justified statements to support conclusions.	Please verify if the student: • knows how to identify trend lines and line of best fit. • knows how to generate justification statements based on data.

* Based on student performance data from the current as well as past tests.
** General recommendations are based on best practices and current research taking into account this student's performance and performance of students at this *level*.

5 of 5

Figure 6. **What to do next (p. 5 of 5).**



Classroom Score Report

Teacher: Ms. Brown
Class: 3rd Period
Grade: 8
Test: 3 of 3
Test Date: 03/30/10

Classroom > Introduction > Appropriate Use AND Inappropriate Use > Performance Summary >

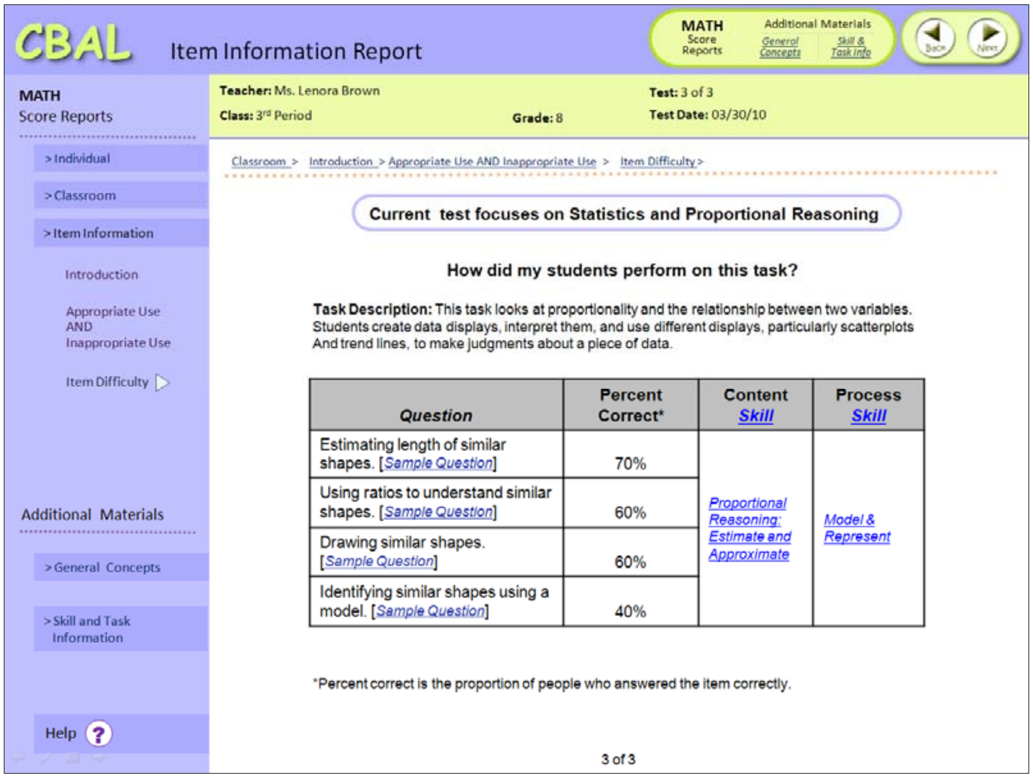
How did my students perform on *Mathematics* overall?

Name	Class Period	*Score* * [0 - 400]	*Performance Level* *
Smith, Pat	1	240	Proficient
Katz, Jonathan	1	90	Not Yet Basic
Bing, Chi	1	380	Advanced
Jackson, Thomas	1	270	Proficient
Gallagher, Walter	1	130	Basic
Denning, Leticia	1	240	Proficient
Hui, Kelly	2	190	Basic
Ghandi, Malia	2	350	Advanced
Diaz, Ana	2	220	Proficient
Chen, Lee	2	170	Basic

• Score and performance level information are based on aggregate student performance data from the current as well as past tests. *Test scores* are not perfectly precise. The "*true score*" lies in an interval whose length depends on the confidence level and the standard error of measurement –*SEM* (**10**).

3 of 3

Figure 7. **Classroom score report for teachers (Math) - performance summary (Pg. 3).**



Teacher: Ms. Lenora Brown

Class: 3rd Period

Grade: 8

Test: 3 of 3

Test Date: 03/30/10

Current test focuses on Statistics and Proportional Reasoning

How did my students perform on this task?

Task Description: This task looks at proportionality and the relationship between two variables. Students create data displays, interpret them, and use different displays, particularly scatterplots And trend lines, to make judgments about a piece of data.

Question	Percent Correct*	Content Skill	Process Skill
Estimating length of similar shapes. [Sample Question]	70%	Proportional Reasoning: Estimate and Approximate	Model & Represent
Using ratios to understand similar shapes. [Sample Question]	60%		
Drawing similar shapes. [Sample Question]	60%		
Identifying similar shapes using a model. [Sample Question]	40%		

*Percent correct is the proportion of people who answered the item correctly.

3 of 3

Figure 8. **Item information report (math) - Item difficulty (p. 3).**

Evaluation

Results of a usability study using a previous version of the score reports with 12 sixth-grade to eighth-grade teachers from schools in NJ and PA showed that teachers reacted positively to interacting with the on-line score reports. However, some of them requested that hardcopy printouts also be made available. In general, the teachers preferred information presented as short, easy-to-read pieces. Long paragraphs were often ignored; after having read 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 seem 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 two levels of report version (current vs. enhanced–additional links to help topics) with two tutorial conditions (tutorial administered vs. 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. Although the total test score did not vary to a statistically significant degree across experimental conditions, the responses to particular questions showed that the teachers (94% or more) could recognize the correct use of each score report. However, a significant proportion of the teachers were willing to consider other uses and purposes that were not valid. For example, 54% of the teachers agreed with the statement “a valid use for the student score report is to place students in advanced or special programs,” and 57% believed that “a valid use for the student score report is to evaluate the current math curriculum.”

Prior versions of the score reports included a list of appropriate uses only. Current versions include both a list of appropriate and inappropriate uses (e.g., see Figure 3). This study also showed that teachers 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, VanWinkle, and Zwick (2010).

Future work includes refining and evaluating the score reports (e.g., improving the wording of statistical information, minimizing the use of technical terms, and exploring alternative graphical representations).

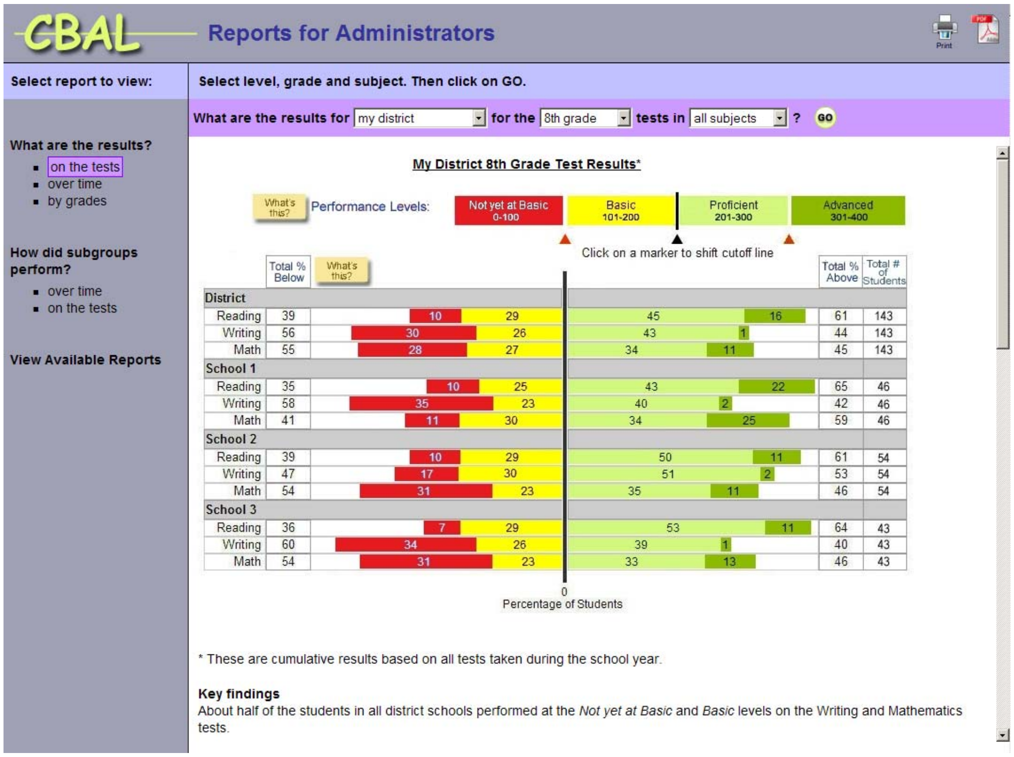
Reports for Administrators

School district administrators experience multiple external demands from various sources forcing them to ignore certain demands, accommodate others, and reinterpret others (Mac Iver & Farley, 2003). According to Honig and Coburn (2008) administrators increasingly face demands to use "evidence" in their decision making. However, due to the complexity of their responsibilities they do not always make decisions based on sound evidence.

A review of the literature identified seven types of responsibilities for administrators (Underwood, Zapata-Rivera, & VanWinkle, 2010): school improvement plans (Honig, 2003; Honig & Coburn, 2008; Miller, 2003; Wayman, Midgley, & Stringfield, 2005), professional development (Brunner et al., 2005; Coburn, Honig, & Stein, 2009; Honig & Coburn, 2008; Mac Iver & Farley, 2003), program selection and evaluation (Brunner et al., 2005; Coburn & Talbert, 2005; Guerard, 2001; Honig, 2003; Honig & Coburn, 2008), curriculum selection (Coburn et al, 2009; Honig & Coburn, 2008; Mac Iver & Farley, 2003), improving student achievement

(Coburn & Talbert, 2005), communication (Chen, Heritage, & Lee, 2005) and staff allocation (Honig & Coburn, 2008).

We have designed a prototype report system taking into account the questions these stakeholders want answered based on their responsibilities. This section describes some of the reports available for these stakeholders. Figure 9 shows a report that is generated after a user decides to view overall results for the tests (selected from the left-hand navigation menu). First, the user makes a selection between overall results or subgroups from the left-hand navigation menu. Next, the user chooses among results for tests, over time, or by grades and then makes selections from the drop-down menu options presented at the top of the screen. In this example, “my district”, “8th grade”, and “all subjects” were selected. Finally, the user clicks on the GO button to generate the score report.



Select level, grade and subject. Then click on GO.

What are the results for my district for the 8th grade tests in all subjects ?

My District 8th Grade Test Results*

Performance Levels: Not yet at Basic 0-100 | Basic 101-200 | Proficient 201-300 | Advanced 301-400

	Total % Below	Not yet at Basic	Basic	Proficient	Advanced	Total % Above	Total # of Students
District							
Reading	39	10	29	45	16	61	143
Writing	56	30	26	43	1	44	143
Math	55	28	27	34	11	45	143
School 1							
Reading	35	10	25	43	22	65	46
Writing	58	35	23	40	2	42	46
Math	41	11	30	34	25	59	46
School 2							
Reading	39	10	29	50	11	61	54
Writing	47	17	30	51	2	53	54
Math	54	31	23	35	11	46	54
School 3							
Reading	36	7	29	53	11	64	43
Writing	60	34	26	39	1	40	43
Math	54	31	23	33	13	46	43

* These are cumulative results based on all tests taken during the school year.

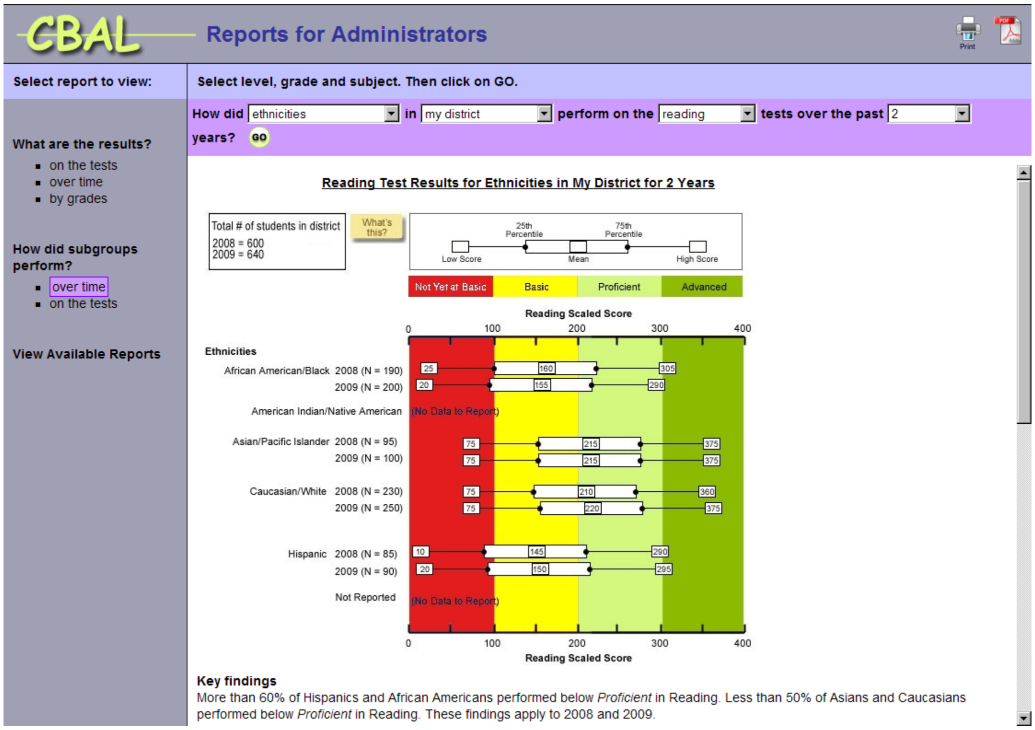
Key findings
About half of the students in all district schools performed at the *Not yet at Basic* and *Basic* levels on the Writing and Mathematics tests.

Figure 9. **Results for my district for the 8th grade tests in all subjects (fragment).**

The icons in the top right-hand corner allow the user to print the report or save the report as a PDF. The performance level legend at the top is interactive and allows the user to click on the marker to shift the cutoff line. Shifting the cutoff line displays the total percent below and above or below a specific performance level. The “What's This” rollovers next to the performance-levels legend and the “Total % Below” column provide users with additional information to help them interpret that area of the score report. Each row provides the percentage of students falling in each of the performance levels, as well as the total number of students. The information in the representation is organized according to district and school, as well as the subject. Additional information appears below the graphical representation. This information includes key findings, a written summary of the main results, a purpose and use section, definitions, interpretations, and links to related reports.

Figure 10 shows a report generated when a user selects subgroups over time from the left-hand navigation menu. The “What's This” rollover provides additional information to help interpret the graph. The total number of students in the district is provided on the left-hand side of the report. The scale is provided at the top and bottom of the graph. The low, mean, and high scores are provided in boxes for each subgroup. Similar to the overall results reports, additional information including key findings, main results, a purpose and use section, definitions, interpretations, and related reports is provided below the graph.

Users can also access the reports by clicking on the “View Available Reports” link on the left-hand navigation menu. This link allows a user to see all of the reports that are available.



CBAL Reports for Administrators

Select report to view:

What are the results?
- on the tests
- over time
- by grades

How did subgroups perform?
- over time
- on the tests

View Available Reports

Select level, grade and subject. Then click on GO.

How did ethnicities in my district perform on the reading tests over the past 2 years?

Reading Test Results for Ethnicities in My District for 2 Years

Total # of students in district
2008 = 600
2009 = 640

Key findings

More than 60% of Hispanics and African Americans performed below *Proficient* in Reading. Less than 50% of Asians and Caucasians performed below *Proficient* in Reading. These findings apply to 2008 and 2009.

Figure 10. **Reading results divided by ethnicities over the past 2 years (fragment).**

Evaluation

A group of seven local administrators participated in a usability study. Overall, we found that the participants found the question-based navigation approach clear and useful, liked the use of “What's This” rollovers and the interactive performance-level legend, found the key findings reasonable, and found the purpose and use section clear. Two out of the seven participants were not able to correctly explain the boxplot. Based on this information, the text provided in the “What's This” rollover was revised to facilitate user comprehension.

A group of three external experts also reviewed the reports and provided recommendations. In general, experts found the reports accessible and easy to understand. Some of the recommendations provided include:

- **Vertical bars.** One of the experts thought it would be easier for administrators to interpret the representations if the bars were vertical instead of horizontal. Experts suggested carrying out a study to explore whether using vertical bars improves teacher understanding of the reports.

- **Color use.** They also wanted color to be used in a more meaningful manner. Specifically, they suggested using different shades of the same color that would become darker as the performance level increased (this change has already been made to the teacher and student reports).

- **Attendance information.** It was also suggested that we include information about attendance. This information is useful to administrators who may want to see how many days of school were missed on average by students who are performing at a certain level.

- **Standard error of difference.** An expert recommended including the standard error of difference when comparing subgroups.

- **Regrouping rows.** Another recommendation was to regroup the rows in the graphical display, so the rows show information groups by subject followed by district and school.

- **Moving number of students column.** Finally, the experts suggested moving the column that displays the total number of students to the left of the representation. Currently, a user must read through the row to find this information and this may be missed.

Future work includes revising the reports based on the feedback gathered, linking the score reports to classroom as well as individual student reports for teachers, and carrying out additional studies exploring how alternate representations influence administrators' understanding of and access to report information.

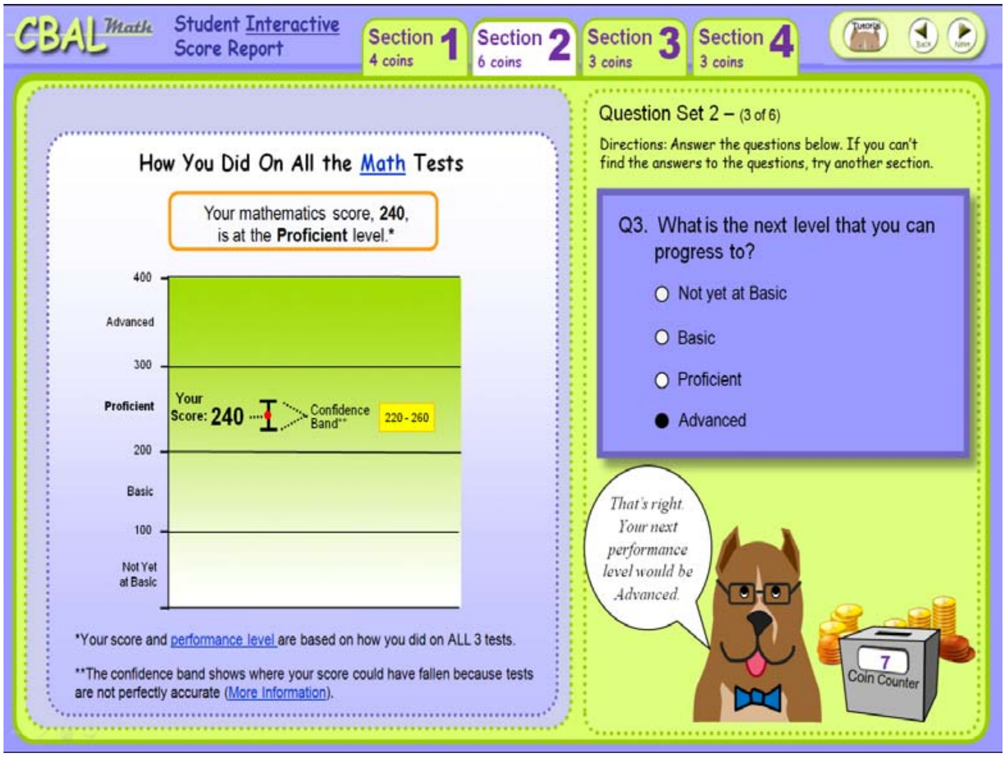
Interactive Score Reports for Students

A review of commercially available score reports showed that most of the student score reports are aimed at parents (Underwood, Reshetar, & Leahy, 2006). Parents want to know the student's overall score, the passing score, or cut scores for different proficiency levels, how the student's score compares to other scores, progress made in different areas, and specific recommendations for helping their children. Although this information is important for both parents and students, in general students have played a passive role in the design of the score reports, which results in score reports that look similar to those score reports provided to teachers, with some modifications (e.g., language employed).

A review of existing score reports for students shows that, currently, student score reports are usually static PDF documents that include technical terms that students do not understand. In addition, score reports are usually available at the end of the academic year, which limits their use for guiding student learning. It is not surprising that students, who are accustomed to highly interactive communication and entertainment tools, find these score reports unattractive, disengaging, and somewhat disconnected from their learning process.

In order to effectively communicate assessment information to students, score reports need to engage them in an activity that encourages them to understand the contents of the score report and use this information to guide their learning process. These new types of score reports should not only communicate assessment information clearly, but also support student motivation and student learning.

Thinking about making students active participants in their learning process and considering their score reporting needs, we have designed a new interactive student score report that implements a guided instructional activity aimed at facilitating student understanding of score report information and improving student engagement. This guided instructional activity consists of using a tabbed menu to navigate through the different sections of the score report and collect coins by correctly answering questions about the content of the report. Coins that students collect are put in a safe, which displays the number of coins that they have collected (see Figure 11). More difficult questions in the report are worth more coins. Students that collect the most coins will earn a spot on a high-score list.



How You Did On All the Math Tests

Your mathematics score, **240**, is at the **Proficient** level.*

*Your score and performance level are based on how you did on ALL 3 tests.

**The confidence band shows where your score could have fallen because tests are not perfectly accurate (More Information).

Question Set 2 – (3 of 6)

Directions: Answer the questions below. If you can't find the answers to the questions, try another section.

Q3. What is the next level that you can progress to?

○ Not yet at Basic

○ Basic

○ Proficient

● Advanced

That's right. Your next performance level would be Advanced.

***Figure 11*. How you did on all the math tests.**

A virtual character guides the students through the score report and provides feedback on their answers. After they have explored all of the sections of the score report, students write about their performance and propose an action plan. Students are given the opportunity to share this action plan with the teacher and/or parents or guardians.

The score report has four sections: identifying information and purpose of the report (Section 1), overall performance results based on the current test as well as past tests (Section 2), performance on the most recent test (Section 3) and a list of areas that should be mastered in order to progress to the next performance level (Section 4). The report also includes a short tutorial.

Evaluation

Results of a usability study with eight local middle school students suggest that students find the activity engaging and the contents of the score report clear. Most of the students were able to explain statistical terms, such as confidence band, with the help of information from the score report. Students made suggestions aimed at improving the look and feel of the score report. Students appreciated the opportunity to share their own improvement plan with the teacher or parents/guardians.

A group of three external experts reviewed the score report and provided general feedback and recommendations for future work. Experts appreciated the effort to design score reports for students, considered the current work innovative, and referred to this work as being on the right path. Experts’ suggestions included: simplifying the amount of information presented to students and carrying out small studies to evaluate the graphical representations, definitions, and feedback available to students in the score report.

Summary and Future Work

This paper reviews relevant literature in the area, presents a framework for designing and evaluating score reports with particular audiences, and describes score reports for three different audiences: teachers, administrators, and students.

Future work also includes: conducting studies aimed at evaluating various graphical representations and other score report information (e.g., definitions, feedback, and interaction aspects), developing and evaluating score reports that are accessible to users with disabilities or those who are English language learners, exploring how information gathered using these reports can be used to guide the development of formative materials for teachers, and further exploring the use of reports as communication tools aimed at supporting sharing of assessment information among teachers, students, parents, and other educational stakeholders.

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: There were 6 grammatical errors in this essay. *Note.* No claim is made about what the student knows or can do, only about the student's response to the test.	Example: John may need to work on grammar, including subject-verb agreement. *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).	Example: *Total Test*: Meets Expectations *Formulate Arguments*: Meets Expectations *Assess Arguments*: Below Expectations *Note.*The claim is about what the student knows and can do.	Example: *Total Test*: 225 *Formulate Arguments*: 230 *Assess Arguments*: 175 *Note.* The claim is about what the student knows and can do.
Evidence requirements	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 student's hypotheses and focused diagnostic measures or teacher judgments.	Data indicating (a) the probability a student's perform-ance 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 scaled 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

References

Alter, A. L., & Oppenheimer, D. M. (2008). Effects of fluency on psychological distance and mental construal (or why New York is a large city, but *New York* is a civilized jungle). *Psychological Science, 19*(2), 161–167.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, *139*(4), 569–576.

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

Axsom, D., Yates, S., & Chaiken, S. (1987). Audience response as a heuristic cue in persuasion. *Journal of Personality and Social Psychology*, *53*(1), 30–40.

Ben-Joseph, E. P., Dowshen, S. A., & Izenberg, N. (2009). Do parents understand growth charts? A national, internet-based survey. *Pediatrics, 124*(4), 1100–1109.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–62). Dordrecht, the Netherlands: Springer.

Bennett, R. E., & Shepherd, M. J. (1982). Basic measurement proficiency of learning disability specialists. *Learning Disability Quarterly, 5*(2), 177–184.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York, NY: Worth Publishers.

Bless, H., & Schwarz, N. (1999). Sufficient and necessary conditions in dual-process models: The case of mood and information processing. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 423–440). New York, NY: Guilford.

Brunner, C., Fasca, C., Heinze, J., Honey, M., Light, D., Mardinach, E., & Fasca, C. (2005). Linking data and learning: The Grow Network study. *Journal of Education for Students Placed at Risk, 10*(3), 241–267.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116–131.

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology, 45*(4), 805–818.

Canham, M., & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction*, *20*(2), 155–166.

Carswell, C. M., & Wickens, C. D. (1987). Information integration and the object display: An interaction of task demands and display superiority. *Ergonomics*, *30*, 511–527.

Casner, S. M., & Larkin, J. H. (1989). *Cognitive efficiency considerations for good graphic design* (Technical Report AIP-81). Retrieved from the Defense Technical Information Center website: http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA218976

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(5), 752–766.

Chandler, P. & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*(4), 293–332.

Chen, E., Heritage, M., & Lee, J. (2005). Identifying and monitoring students' learning needs with technology. *Journal of Education for Students Placed at Risk*, *10*(3), 309–332.

Chua, H. F., Liberzon, I., Welsh, R. C., & Strecher, V. J. (2009). Neural correlates of message tailoring and self-relatedness in smoking cessation programming. *Biological Psychiatry, 65*(2), 165–168.

Chua, H. F., Yates, J. F., & Shah, P. *(2006). Risk avoidance: Pictures versus numbers.* Memory & Cognition, 34*(2), 399–410.*

Cleveland, W. S. (1985). *The elements of graphing data*. Monterey, CA: Wadsworth Advanced Books and Software.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*(387), 531–554.

Coburn, C. E., & Talbert, J. E. (2005, April). Conceptions of evidence use in school districts: Mapping the terrain. *Paper presented at the annual meeting of the American Educational Research Association*, Montreal, Canada.

Coburn, C. E., Honig, M. I., & Stein, M. K. (2009). What's the evidence on districts' use of evidence? In J. Bransford, D. J. Stipek, N. J. Vye, L. Gomez, & D. Lam (Eds.), *The role of research in educational improvement* (pp. 67–88). Cambridge, MA: Harvard Education Press.

Deng, N., & Yoo, H. (2009) *Resources for Reporting Test Scores: A Bibliography for the Assessment Community*. Prepared for the National Council on Measurement in Education. Retrieved from https://ncme.org/resources/blblio1/NCME.Bibliography-5-6-09_score_reporting.pdf

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. (2010). Fortune favors the **bold** (*and the Italicized*): Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 111–115. doi:10.1016/j.cognition.2010.09.012

English, L. D. (1997). *Mathematical reasoning: Analogies, metaphors, and images.* Mahwah, NJ: Erlbaum.

Fagerlin, A., Ubel, P., Smith, D., & Zikmund-Fisher, B. (2007). Making numbers matter: Present and future research in risk communication. *American Journal of Health Behavior*, *31*(1), S47–S56.

Fagerlin, A., Wang, C., & Ubel, P. (2005). Reducing the influence of anecdotal information on people's health care decisions: Is a picture worth 1,000 statistics? *Medical Decision Making*, *25*(4), 398–405.

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 easily understood format*. Council of Chief State School Officers. Washington, DC.

Freedman, D. Pisani, R., Purves, R., & Adhikari, A. (1991). *Statistics* (2nd ed.). New York, NY: W. W. Norton.

Gillan, D. J. & Richman, E. H. (1994). Minimalism and the syntax of graphs. *Human Factors*, *36*(4), 619– 644.

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*(2), 145–220.

Guerard, E. B. (2001, May 17). School leaders learn to make data-driven decisions. *eSchool News*. Retrieved from http://www.sanbenito.k12.tx.us/departments/technology/pdf/esnbestpractices.pdf

Guthrie, J. T., Weber, S. & Kimmerly, N. (1993). Searching documents: Cognitive processes and deficits in understanding graphs, tables, and illustrations. *Contemporary Educational Psychology, 18*, 186–221.

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., & Zenisky, A. (2010, November). *Improvements to student score reporting: steps and more use of suitable methodologies.* Presentation at the ETS Conference on Score Reporting, November 4–5, 2010.

Hattie, J. (2009). *Visibly learning from reports: The validity of score reports*. Retrieved from http://www.oerj.org/View?action=viewPDF&paper=6

Hegarty M. (2004). Diagrams in the mind and in the world: Relations between internal and external visualizations. In A. Blackwell, K. Mariott & A. Shimojima (Eds.). *Diagrammatic representation and inference. Lecture notes in artificial intelligence 2980* (pp. 1–13). Berlin, Germany: Springer-Verlag.

Hegarty, M., Kriz, S. & Cate, C. (2003). The roles of mental animations and external animations in understanding mechanical systems. *Cognition & Instruction, 21*(4), 209–249.

Honig, M. I. (2003). Building policy from practice: District central office administrators' roles and capacity for implementing collaborative education policy. *Educational Administration Quarterly, 39*(3), 292–338.

Honig, M. I., & Coburn, C. E. (2008). Evidence-based decision-making in school district central offices: Toward a policy and research agenda. *Educational Policy, 22*, 578–608.

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,* 16–18.

Kosslyn, S. M. (1994). *Elements of graph design.* New York, NY: Freeman.

Kriz, S., & Hegarty, M. (2007). Top-down and bottom-up influences on learning from animations. *International Journal of Human-Computer Studies, 65,* 911–930.

Larkin, J. & Simon, H. (1987) Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*(1), 65–100.

Levy, E., Zacks, J., Tversky, B. & Schiano, D. (1996). Gratuitous graphics: Putting preferences in perspective. Human factors in computing systems: Conference proceedings (pp. 42–49). New York, NY: ACM.

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.

Mac Iver, M. A., & Farley, E. (2003). *Bringing the district back in: The role of the central office in improving instruction and student achievement* (CRESPAR Report No. 65). Baltimore, MD: Johns Hopkins University.

Mayer, R. E. (2001). *Multimedia Learning*. Cambridge, UK: Cambridge University Press.

Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction, 13,* 125–139.

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. (2002). Animation as an aid to multimedia learning. *Educational Psychology Review, 14,* 87–99.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38* (1), 43–52.

Mayer, R.E. (2001). *Multimedia learning.* Cambridge, England: Cambridge University Press.

Miller, N. (2003, February). Best practices. *Super Tech News, 2*(1). Retrieved from http://www.blegroup.com/supertechnews/feb03.htm#best.

Moreno, R., Mayer, R., Spires, H., & Lester, J. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction, 19,* 177–213.

Narayanan, N. H., & Hegarty, M. (2002). Multimedia design for communication of dynamic information. *International Journal of Human-Computer Studies, 57*, 279–315.

National Council on Measurement in Education (NCME). (1995). *Code of professional responsibilities in educational measurement.* Retrieved from http://www.ncme.org/about/docs/prof_respons.doc

Neisser, U. (1963). Decision time without reaction time: Experiments in visual scanning. *American Journal of Psychology, 76*, 376–385.

Nielsen, J. (1994). *Usability engineering*. San Francisco, CA: Morgan Kaufmann.

Noar, S. M., Benac, C. N., & Harris, M. S. (2007). Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychological Bulletin, 133*(4), 673–693.

Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2007). Preference fluency and consumer choice. *Journal of Marketing Research, 44*(3), 347–356.

Nussbaum, E. M. (2005). The effect of goal instructions and need for cognition on interactive argumentation. *Contemporary Educational Psychology,* 30(3), 286–313.

Olzak, L. A., & Thomas, J. P. (1986). Seeing spatial displays. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.). *Handbook of perception and performance: Sensory processes and perception* (pp. 717–756). New York, NY: Wiley.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). New York, NY: Academic Press.

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73–126). Hillsdale, NJ: Lawrence Erlbaum Associates.

Reber, R., & Zupanek, N. (2002). Effects of processing fluency on estimates of probability and frequency. In P. Sedlmeier & T. Betsch (Eds.), *Frequency processing and cognition* (pp. 175–188). Oxford, UK: Oxford University Press.

Roberts, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice, 29*(3), 25–38.

Rodriguez, F., Shah, P., & Ng, A. (March, 2010). *What reasoning strategies do novice college students use to critically evaluate scientific evidence?* Paper presented at the 13th biennial meeting of the Society for Research on Adolescence, Philadelphia, PA.

See, Y. H., Petty, R. E., & Evans, L. M. (2009). The impact of perceived message complexity and need for cognition on information processing and attitudes. *Journal of Research in Personality, 43*(5), 880–889.

Seife, C. (2010). *Proofiness: The dark arts of mathematical deception.* New York, NY: Viking.

Shah, P. Freedman, E., & Miyake, A. (2011). *Are labels really better than legends?* Manuscript in review.

Shah, P., & Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General, 124*, 43–61.

Shah, P., & Freedman, E. (2009). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science, 3*(3), 560–578

Shah, P., Freedman, E., & Vekiri, I. (2005). The comprehension of quantitative information in graphical displays. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 426–476). New York, NY: Cambridge University Press.

Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology, 91*(4), 690–702.

Simmons, J. P., & Nelson, L. D. (2006a). Intuitive confidence: Choosing between intuitive and nonintuitive alternatives. *Journal of Experimental Psychology: General, 135*(3), 409–428.

Simmons, J. P., & Nelson, L. D. (2006b). *Intuitive confidence and the prominence effect: When consumer choices are sensitive to matching prices.* Manuscript in preparation.

Sklar, J., & Zwick, R. (2009). Multimedia presentations in educational measurement and statistics. *Journal of Statistics Education.* http://www.amstat.org/publications/jse/v17n3/sklar.html.

Smith, L. D., Best, L. A., Stubbs, D. A., Johnston, J., & Archibald, A. B. (2000). Scientific graphs and the hierarchy of the sciences: A Latourian survey of inscription practices. *Social Studies of Science, 30,* 73–94.

Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology, 5*, 61–77.

Stiggins, R., & Herrick, M. (2007). A status report on teacher preparation in classroom assessment. Unpublished manuscript.

Sweller, J. (1999). *Instructional design in technical areas*. Camberwell, Australia: ACER Press.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin, 121*(3), 371–394.

Tetlock, P. E., & Lerner, J. S. (1999). The social contingency model: Identifying empirical and normative boundary conditions for the error- and-bias portrait of human nature. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 571–585). New York, NY: Guilford 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.

Twing, J. (March, 2008). *Score reporting, off-the-shelf assessments and NCLB: Truly an unholy trinity*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Underwood J. S., Reshetar, R., & Leahy, S. (2006). *Score report design heuristics.* Unpublished report.

Underwood J. S., Zapata-Rivera D., & VanWinkle, W. (2007). *Growing pains: Teachers using and learning to use IDMS.* (ETS Research Rep. No. RM-08-07). Princeton, NJ: ETS.

Underwood, J. S., Zapata-Rivera, D., & VanWinkle, W. (2010). *An evidence-centered approach to using assessment data for policymakers.* (ETS Research Rep. No. RR-10-03*)*. Princeton, NJ: ETS.

Utts, J. M., & Heckard, R. F. (2004). *Mind on Statistics* (2nd ed.). Belmont, CA: Brooks/Cole.

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.

Wainer, H. (2009). *Picturing the Uncertain World: How to Understand, Communicate and Control Uncertainty through Graphical Display*. Princeton, NJ: Princeton University Press.

Wayman, J. C., Midgley, S., & Stringfield, S. (2005, April). *Collaborative teams to support data-based decision making and instructional improvement*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Zacks, J., & Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory and Cognition, 27*(6), 1073–1079.

Zapata-Rivera D., & VanWinkle, W. (2010). *A research-based approach to designing and evaluating score reports for teachers.* ETS Research Memorandum No.RM-10-01. Princeton, NJ: ETS.

Zapata-Rivera, D. & VanWinkle, W., & Zwick, R. (2010, April). *Exploring effective communication and appropriate use of assessment results through teacher score reports.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Denver, CO.

Zikmund-Fisher, B., Fagerlin, A., & Ubel, P. (2008). Improving understanding of adjuvant therapy options by using simpler risk graphics. *Cancer, 113*(12), 3382–2290.

Zwick, R., Sklar, J., Wakefield, G., Hamilton, C., Norman, A., & Folsom, D. (2008). Instructional tools in educational measurement and statistics (ITEMS) for school personnel. *Educational Measurement: Issues and Practice, 27*, 14–27.

Notes

[1] This work was conducted when the first author was at University of California, Santa Barbara. This project was funded by the National Science Foundation (#0352519). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

[2] The two articles and all ITEMS videos, quizzes, and supplementary materials are available at http://items.education.ucsb.edu