



Research Memorandum
ETS RM-13-04

Exploring Teachers' Understanding of Graphical Representations of Group Performance

Diego Zapata-Rivera

Margaret Vezzu

Waverely VanWinkle

May 2013

ETS Research Memorandum Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Research Scientist

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Research Scientist

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Gary Ockey
Research Scientist

Donald Powers
Managing Principal Research Scientist

Frank Rijmen
Principal Research Scientist

John Sabatini
Managing Principal Research Scientist

Matthias von Davier
Director, Research

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ruth Greenwood
Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

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

Exploring Teachers' Understanding of Graphical Representations of Group Performance

Diego Zapata-Rivera, Margaret Vezzu, and Waverley VanWinkle
ETS, Princeton, New Jersey

May 2013

Find other ETS-published reports by searching the ETS
ReSEARCHER database at <http://search.ets.org/researcher/>

To obtain a copy of an ETS research report, please visit
<http://www.ets.org/research/contact.html>

Action Editor: Donald E. Powers

Reviewers: Irvin R. Katz and Madeleine Keehner

Copyright © 2013 by Educational Testing Service. All rights reserved.

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



Abstract

Every year in the United States, millions of score reports are produced and delivered to teachers, students, parents, school administrators, and policymakers. Research shows that many educators have trouble understanding and making appropriate use of score reports. This paper evaluates three graphical representations of test score distribution. Results of a usability study and a study evaluating these three graphical representations are presented and discussed.

Key words: graphical representations, distribution of scores, design and evaluation of score reports, score reports for teachers

Acknowledgments

We would like to acknowledge Janet Stumper and Tom Florek for their excellent work on the project. We also want to thank Rebecca Zwick for her useful comments and suggestions. We would like to extend our gratitude to the teachers who participated in our studies. Finally, we thank Madeleine Keehner, Irv Katz, and Don Powers for providing useful comments on an earlier version of this report.

Table of Contents

	Page
Background.....	1
Methodology.....	2
Participants	4
Administration	4
Instruments	5
Usability Study	5
Results.....	6
Comprehension Results	6
Item Analysis Results	6
Preference Results	7
Background Variables and Comprehension Results.....	9
Comfort With Statistics and Preference Results.....	10
Analysis of Explanations	10
Summary and Discussion.....	16
Future Work.....	17
References.....	18
Notes	20
Appendix A. Comprehension Questionnaire	21
Appendix B. Preference Questionnaire.....	29

List of Tables

	Page
Table 1. Participant Demographic Characteristics.....	4
Table 2. Mean Comprehension Scores, Standard Deviations, and Sample Sizes for the Three Conditions	6
Table 3. Response to Question 1a, “Which of These Three Representations Do You Prefer?”..	9
Table 4. Response to Question 2a, “Which of These Three Versions Helps You Understand the Distribution Better?”	8
Table 5. Response to Question 3a, “Which Version Makes It Easier for You to Identify the Number of Students Whose Score Fell Within a Specific Range?”	9
Table 6. Response to Question 4a, “Which Version Do You Think Other Teachers Would Be Able to Understand the Best?”	9
Table 7. Comprehension Score for Participants With Low and High Comfort With Statistical Terms.....	10
Table 8. Comfort With Statistical Terms and Concepts and Preference.....	10
Table 9. Explanation Categories and Counts Associated With Question 1a, “Which of These Three Representations Do You Prefer?”	11
Table 10. Explanation Categories and Counts for Question 2a, “Which of These Three Versions Helps You Understand the Distribution Better?”	12
Table 11. Explanation Categories and Counts for Question 3a, “Which Version Makes It Easier for You to Identify the Number of Students Whose Score Fell Within a Specific Range?”	14
Table 12. Categories and Counts for Question 4a, “Which Version Do You Think Other Teachers Would Be Able to Understand the Best?”	15
Table 13. Categories and Counts for Question 5, “How Could These Score Reports Be Made More Effective?”	16

List of Figures

	Page
Figure 1. Box-and-whisker plot.....	3
Figure 2. Stacked report icons showing where the scores fall.....	3
Figure 3. Curve showing the distribution of scores.....	4

Research on score reports have indicated that teachers, policymakers, and students have trouble understanding the terminology and graphical displays used to communicate assessment results (e.g. Hambleton & Slater, 1997; Lukin, Bandalos, Eckhout, & Mickelson, 2004; Zapata-Rivera, VanWinkle, & Zwick, 2010; Zwick et al., 2008). Several strategies for dealing with this problem have been explored, including (a) proposed heuristics and best practices based on research findings from fields such as information visualization and graphical representations of quantitative data (e.g., Goodman & Hambleton, 2004; Zenisky & Hambleton, 2012); (b) taking into account the characteristics of the target audience (e.g., knowledge, preferences, goals) when designing score reports and providing interpretive materials (e.g., Roberts & Gierl, 2010; Zapata-Rivera, 2011); (c) evaluating score reports with experts and the intended audience before the score reports are deployed (e.g., VanWinkle, Vezzu, & Zapata-Rivera, 2011; Vezzu, VanWinkle, & Zapata-Rivera, 2012); and (d) educating the intended audience on graphical representations and educational assessment issues (e.g., professional development courses and instructional materials such as online help and tutorials; Zapata-Rivera, VanWinkle, & Zwick, 2010; Zwick et al., 2008).

A particular area of interest is exploring how graphical representations of assessment information can help teachers understand and make good use of assessment results. Graphical representations of group performance results (distribution of scores) are usually included in score reports for teachers and administrators. This paper evaluates three graphical representations of distribution of scores. We report on the results of a study aimed at exploring the effectiveness of these three graphical representations at communicating distribution of scores results to teachers. This work has been done in the context of ETS's Cognitively Based Assessment of, for, and as Learning (*CBAL*TM) research initiative (Bennett & Gitomer, 2009).¹

Background

Research on how students develop an understanding of the concept of a frequency distribution shows that students usually go from reasoning at the individual-value level to forming an informal understanding of the characteristics of a distribution (e.g., using less precise terms such as *bumps* in relation to the shape of a distribution). With additional practice, students also learn to reason at the conceptual level, understanding particular characteristics of a distribution (e.g., center, spread, density, skewness) and reasoning about particular data values

(e.g., outliers). Experts can easily combine these modes of reasoning (Bakker & Gravemeijer, 2005).

Interactive computer-based tools have been designed to guide students' understanding of the concept of a distribution. Bakker and Gravemeijer (2005) reported that using Minitools (Cobb, 1999), students acquire an informal understanding of the concept of distribution and transition into the use of more precise definitions. Minitools employs different types of graphical representations (e.g., dot plots, bar plots [vertical and horizontal], box-and-whisker plots). These graphical representations offer different degrees of detail supporting reasoning at the individual-value or conceptual levels.

Tools for teaching the concept of distribution often include interactive applications where users can see individual values or complete distributions. Some of these tools include: TinkerPlots (Steinke, 2005), Tabletop Jr. and Tabletop (Bagnall, 1994), and Fathom Dynamic Data software (Key Curriculum Press, 2011).

A common graphical representation used to summarize group performance information (e.g., distribution of scores) is the box-and-whisker plot (Spear, 1952; Tukey, 1977). This graphical representation is usually found in score reports for teachers and policymakers. Previous work on evaluating reports for policymakers suggested that some participants were confused by this representation. For example, even after reviewing the box-and-whisker representation of the distribution of scores, some participants still wanted to see the distribution of scores (VanWinkle, Vezzu, & Zapata-Rivera, 2011). This seems to indicate that some participants are not familiar with this type of representation and that a representation that provides more details (e.g., a representation that shows the shape of the distribution) could possibly be more appropriate for this audience.

Methodology

We have designed three graphical representations for a score distribution showing three different levels of abstraction (see Figures 1–3). Figure 1 shows a score report that uses the traditional box-and-whisker plot. Figure 2 shows where the scores fall by using stacked score report icons. Figure 3 makes use of a curve to show the distribution of scores. These representations are usually used to teach basic statistics.

The traditional box-and-whisker representation, which is the one that we currently use, was included as our control condition. The other two alternative representations (stacked report

icons and curve) were designed to be as parallel as possible (i.e., to convey the same information). Each alternative representation was assigned to one experimental condition.

The purpose of the study was to evaluate the effectiveness of these graphical representations at communicating the concept of frequency distribution (distribution of scores) to teachers.

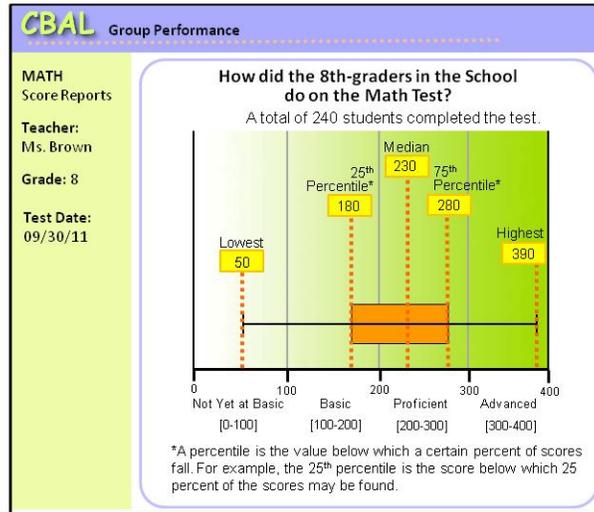


Figure 1. Box-and-whisker plot.²

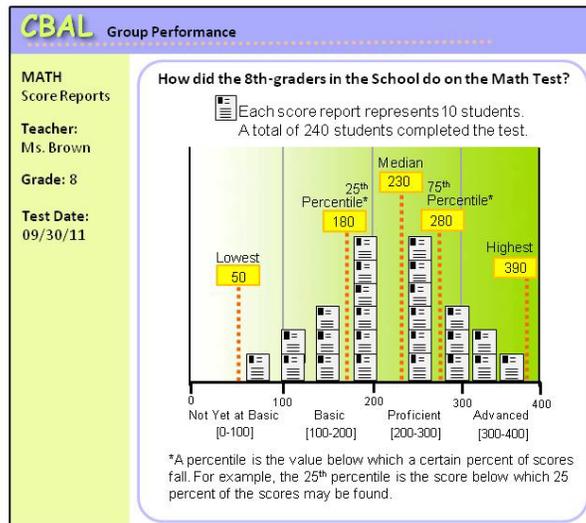


Figure 2. Stacked report icons showing where the scores fall.

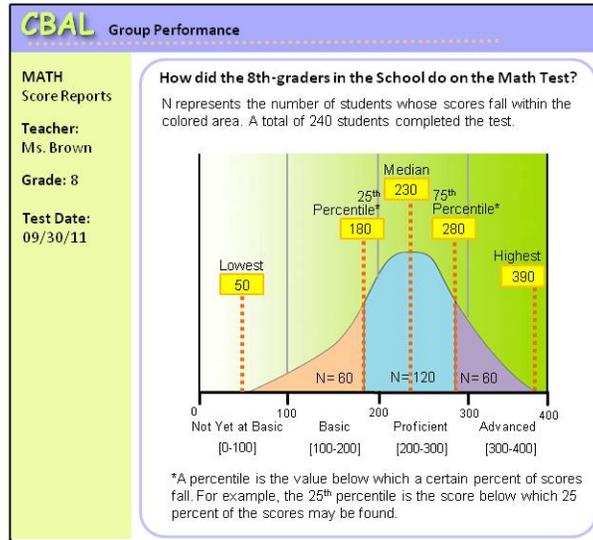


Figure 3. Curve showing the distribution of scores.

Participants

A total of 125 middle school teachers were randomly assigned to one of three conditions. The study lasted for about an hour. Participants received a \$50 gift certificate for their participation. Table 1 shows the demographic characteristics of the participant sample.

Table 1

Participant Demographic Characteristics

Race/ethnicity	Percentage
White	78.4%
Black or African American	13.6%
Asian	1.6%
Native Hawaiian or other Pacific Islander	0
Mexican, Mexican American, or Chicano	0
Puerto Rican	0
Other Hispanic, or Latin American	1.6%
Blank, other, or “decline to state”	4.8%

Note. The percentage of women was 89.6%.

Administration

The study was administered online via the Internet. Teachers were free to choose the location for participating in the study, such as school or home.

Instruments

Three data collection measures were administered:

- A background questionnaire that included demographic information and questions about respondents' background and knowledge in measurement, statistics, and score interpretation.
- A comprehension questionnaire that consisted of 14 multiple-choice questions that assessed the participant's comprehension of the score report. For example, "How many students scored between 50 and 180?" (a) 60; (b) 130; (c) 180; (d) 25 (see Appendix A, Comprehension Questionnaire). These questions were developed to cover many aspects of the score distribution including percentiles, mean, lowest, and highest values. Each question was accompanied by one graphical representation according to the assigned condition.
- A preference questionnaire that included questions where participants could explore the three representations and rate them in terms of usefulness, comprehensibility, and attractiveness. These questions included: "Which of these three representations do you prefer? And why?", "Which of these three versions helps you understand the distribution better? And why?", and "Which version makes it easier for you to identify the number of students whose score fell within a specific range? And why?" (see Appendix B, Preference Questionnaire).

Usability Study

To help ensure the quality of the instruments, a usability study was conducted with six local middle-school teachers. The goal of this study was to identify major accessibility, readability, and navigation problems of the questionnaires before the study. Participants were asked to think aloud while interacting with the reports and responding to the questionnaires. Observers took notes and asked clarification questions. Results from the usability study were transcribed, summarized, and analyzed. Changes made to the graphical representations and questionnaires based on the feedback received from the teachers included: (a) red lines indicating the lowest, the highest, and 25th, 50th, and 75th percentiles were highlighted and extended to the *x*-axis; (b) numerical ranges for performance levels were added; (c) wording of some comprehension questions was edited to avoid misinterpretations; (d) some distracters were

modified; and (e) messages and controls were added to the delivery system to facilitate navigation (e.g., teachers were informed that they had to look at each of the three representations before answering each preference question).

Results

Comprehension Results

Table 2 shows the mean comprehension scores, standard deviations, and sample sizes for the three conditions.

Table 2

Mean Comprehension Scores, Standard Deviations, and Sample Sizes for the Three Conditions

Statistics	Box-and-whisker	Stacked report icons	Curve
Mean score	11.2	12.8	12.8
Standard deviation	3.2	1.3	1.4
Sample size	40	47	38

Note. Maximum possible comprehension score is 14. Teachers were randomly assigned to one of the three conditions.

A Levene’s statistic of 18.58 ($p < 0.05$) indicated that the variability in the three conditions (mean scores) was significantly different. Accordingly, Welch’s ANOVA was run with the post hoc Tamhane’s T2 test. The Welch’s ANOVA showed a significant difference ($p < 0.05$) among groups (Welch = 4.5, $p = 0.014$). Tamhane’s T2 test showed significant differences ($p < 0.05$) between box-and-whisker and stacked report icons ($p = 0.015$) and box-and-whisker and curve ($p = 0.018$). There was no significant difference between the stacked report icons and curve.

Using Cohen’s d based on sample size (Hedges adjustment), relatively large size effects were obtained ($d = 0.65$, Hedges pooled standard deviation = 2.37, when comparing box-and-whisker with stacked report icons and $d = 0.64$, Hedges pooled standard deviation = 2.5, when comparing box-and-whisker with curve).

Item Analysis Results

Item difficulty information can be used to determine which question(s) were more difficult for participants in a particular condition. Item difficulty values (p values³) ranged from

0.68 to 1.0. By setting the cut-off value to 0.8 (p value ≤ 0.8) several questions can be highlighted per condition. For example, Questions 1–4, 6–8, and 10–11 for the box-and-whiskers condition; Question 2 for the stacked report icons condition; and Questions 2 and 8 for the curve condition.

In general, participants assigned to the box-and-whiskers found it more difficult to answer questions that required finding the number of students whose score fell within a particular range because they had to calculate the response by using the total number of students who took the test and percentile/median information.

Question 2 (p values: 0.75–box-and-whiskers, 0.72–stacked report icons, and 0.79–curve) dealt with the number of students whose score fell within a particular performance level. Although the current representations can be used to answer this question, these representations do not clearly show this information. A different representation showing the distribution of scores across performance levels can be designed to support these types of questions.

Question 8 (p values: 0.75–box-and-whiskers, 0.94–stacked report icons, and 0.79–curve) and Question 11 (p values: 0.78–box-and-whiskers, 0.94–stacked report icons, and 0.82–curve) dealt with the number of students within a region involving the mean as one of the bounds. The stacked report icons representation facilitated this type of query by clearly showing report icons below and above the mean.

Question 1 (p values: 0.78–box-and-whiskers, 0.85–stacked report icons, and 0.95–curve) and Question 3 (p values: 0.68–box-and-whiskers, 0.85–stacked report icons, and 0.92–curve) required finding the number of students who scored within two particular scores. These values were easier to find using the curve representation.

Finally, Question 4 (p values: 0.68–box-and-whiskers, 0.96–stacked report icons, and 0.89–curve) asked for the number of students whose score fell within the 25th and 75th percentile. Although the curve representation includes this value, it could be possible that the dotted line representing the mean may have caused some confusion. Participants in the stacked report icons condition just needed to count the report icons.

Preference Results

Participants were asked to answer four preference questions after completing the comprehension questionnaire. Participants were required to look at all of the representations before answering each question and also provided an explanation for their selection in each case.

Tables 3–6 show the percentage of participants who preferred each representation per condition. In general, participants in all conditions preferred the stacked report icons and the curve conditions.

For Question 1a, “Which of these three representations do you prefer?” (Table 3), participants assigned to the box-and-whisker condition slightly preferred the stacked report icons over the curve. Participants assigned to the stacked report icons and the curve conditions preferred the condition they were assigned to in the first place.

For Question 2a, “Which of these three versions helps you understand the distribution better?” Question 3a, “Which version makes it easier for you to identify the number of students whose score fell within a specific range?” and Question 4a, “Which version do you think other teachers would be able to understand the best?” (Tables 4–6), the preferences follow a similar pattern: Participants assigned to the box-and-whisker condition and the curve condition preferred the curve followed by the stacked report icons, while participants assigned to the stacked report icons condition preferred the stacked report icons followed by the curve.

Table 3

Response to Question 1a, “Which of These Three Representations Do You Prefer?”

Condition	Preference: box-and- whisker	Preference: stacked report icons	Preference: curve
Box-and-whisker	12.5%	45.0%	42.5%
Stacked report icons	8.5%	53.2%	38.3%
Curve	13.2%	21.1%	65.8%
Average	11.4%	39.8%	48.9%

Table 4

Response to Question 2a, “Which of These Three Versions Helps You Understand the Distribution Better?”

Condition	Preference: box-and- whisker	Preference: stacked report icons	Preference: curve
Box-and-whisker	10.0%	37.5%	52.5%
Stacked report icons	8.5%	57.5%	34.0%
Curve	18.4%	18.4%	63.2%
Average	12.3%	37.8%	49.9%

Table 5

Response to Question 3a, “Which Version Makes It Easier for You to Identify the Number of Students Whose Score Fell Within a Specific Range?”

Condition	Preference: box-and- whisker	Preference: stacked report icons	Preference: curve
Box-and-whisker	5.0%	32.5%	62.5%
Stacked report icons	6.4%	55.3%	38.3%
Curve	0.0%	26.3%	73.7%
Average	3.8%	38.0%	58.2%

Table 6

Response to Question 4a, “Which Version Do You Think Other Teachers Would Be Able to Understand the Best?”

Condition	Preference: box-and- whisker	Preference: stacked report icons	Preference: curve
Box-and-whisker	2.5%	37.5%	60.0%
Stacked report icons	4.3%	61.7%	34.0%
Curve	13.2%	29.0%	57.9%
Average	6.6%	42.7%	50.7%

Background Variables and Comprehension Results

We examined the relationship between total comprehension scores and various background self-reported variables including: comfort level with statistical terms and concepts; comfort level with computers; number of years teaching; number of undergraduate courses taken in statistics, educational measurement, or psychological measurement; number of graduate courses taken in statistics; educational measurement or psychological measurement; and number of professional development sessions taken in statistics, educational measurement, or psychological measurement.

Among these analyses, only comfort level with statistical terms and concepts was significant, $F(1, 86) = 17.5, p < 0.001$. The corresponding effect size was $d = 0.79$, Hedges pooled standard deviation = 2.13. Comfort with statistical terms and concepts was coded as *low comfort* and *high comfort*. Low comfort participants were those who answered “not at all comfortable” ($N = 5$) or “somewhat comfortable” ($N = 52$) to the question, “How comfortable

are you with statistical terms and concepts (e.g., mean, median, mode, and variance)?” High comfort participants were those who answered “comfortable” ($N = 68$). Table 7 shows descriptive statistics for this variable.

Table 7

Comprehension Score for Participants With Low and High Comfort With Statistical Terms

Comfort level	Mean	SD	N
Low	11.4	2.5	57
High	13.1	1.7	68

Comfort With Statistics and Preference Results

Although both low comfort and high comfort participants seemed to prefer the stacked report icons and the curve conditions over the box-and-whisker condition (see Table 8), low comfort participants slightly preferred the stacked report icons while high comfort participants preferred the curved condition. This seems to indicate that participants that are more familiar with statistical concepts also preferred more abstract representations.

Table 8

Comfort With Statistical Terms and Concepts and Preference

Preferred representation	Comfort level: low	Comfort level: high
Box-and-whisker	4	10
Stacked report icons	31	20
Curve	22	38
Total	57	68

Analysis of Explanations

The explanations provided by the participants about their preferences were coded by two raters individually on previously defined general categories. Cases of disagreement were discussed and consensus was achieved.

Table 9 shows the number of occurrences on each of the explanation categories associated with Question 1a, “Which of these three representations do you prefer?” Layout issues and amount of detail provided were the most cited reasons for preferring a representation.

Some of the explanations provided include “More familiar with this style [curve]” (familiarity); “I prefer this version [stacked report icons] because it is very clear visually. It tells

me exactly how many students scored at each particular level. I was able to see what student's scores were, not just percentile rankings” (provides more details and layout issues); “[curve] It seems to be the easiest to read. You can tell the number of students just by looking at it” (provides more details and layout issues); “Color coding helps to make this [curve] more clear” (layout issues); and “[stacked report icons] I like to see the symbol for 10 students” (likes bars/report icon/paper symbol).

Table 9

Explanation Categories and Counts Associated With Question 1a, “Which of These Three Representations Do You Prefer?”

Explanation category	Preferred: box-and- whisker	Preferred: stacked report icons	Preferred: curve	Total
Familiarity	3	1	8	12
Provides more details: number of students in each range (performance level/percentile), easier to see distribution, and easier to compare individuals with the class	4	26	27	57
Provides a general overview	0	5	2	7
Layout issues: use of color, easier to read, user friendly, clearer, visually attractive, clear text explanations, and ordered	8	29	43	80
Likes box-and-whisker representation	5	1	0	6
Likes bars/report icon/paper symbol	0	7	0	7
Likes the curve	0	0	14	14
N/A, Other	0	1	0	1

Note. An explanation can belong to more than one category.

Table 10 shows the number of occurrences on each of the explanation categories associated with Question 2a, “Which of these three versions helps you understand the distribution better?” Participants felt that layout issues and additional details such as the number

of students in each range and showing the shape of the distribution helped them understand the distribution.

Table 10

Explanation Categories and Counts for Question 2a, “Which of These Three Versions Helps You Understand the Distribution Better?”

Explanation category	Preferred: box-and- whisker	Preferred: stacked report icons	Preferred: curve	Total
Familiarity	2	0	5	7
Shows the highest and the lowest scores	2	0	0	2
Shows number of students in each range (performance level/percentile)/no need to count	1	16	19	36
Shows symbol per each 10 students/I can count	0	15	0	15
Provides a general overview/less details	1	0	1	2
Layout issues: use of color, easier to read, user friendly, clearer, visually attractive, clear text explanations, and ordered	4	17	36	57
Shows the shape of the distribution/easier to see how the data spreads out	4	4	17	25
Likes box-and-whisker representation	1	0	0	1
Likes bars/report icon/paper symbol	0	7	0	7
Likes the curve	0	0	3	3
N/A, Other	1	6	2	9

Note. An explanation can belong to more than one category.

Sample explanations include “This report [stacked report icons] shows exactly how many students scored at each level” (shows number of students in each range); “I am used to the bell curve and the numbers are right there for me. With all the labeling and color coding this is the best” (familiarity, shows number of students in each range, and layout issues); “This one [curve] gives a number of students in each percentile ranking” (shows number of students in each range);

“The symbol that represents 10 students helps you see exactly where the students’ scores were at” (shows symbol per each 10 students); “The bar graphs are clear and concise” (layout issues); “It is more visually appealing [stacked report icons]” (layout issues); “I think that this report [stacked report icons] does represent distribution better...more visual” (shows the shape of the distribution); and “The curve shows the distribution easily. Most of the scores are grouped around the median” (shows the shape of the distribution).

Table 11 shows the number of occurrences on each of the explanation categories associated with Question 3a, “Which version makes it easier for you to identify the number of students whose score fell within a specific range?” Showing a symbol per each 10 students (stacked report icons) and showing the number of students in each range (curve) and layout issues were the reasons participants provided that helped them identify the number of students whose score fell within a specific range.

Some of the explanations provided include “I like the key and knowing that each paper is equal to 10 students” (shows symbol per each 10 students); “I like the symbol - it's easy to count tens” (shows symbol per each 10 students); “The colors as well as the numeric representation allows me to identify the number of students within a specific range [curve]” (shows number of students in each range); “It gives you the exact number of students in each category [curve]” (shows number of students in each range); “The division by color helps me quickly identify the different areas in scoring [curve]” (layout issues); and “The labels help you to see the number of students better [stacked report icons]” (layout issues).

Table 12 shows the number of occurrences for each of the explanation categories associated with Question 4a, “Which version do you think other teachers would be able to understand the best?” Teachers felt that other teachers would be better able to understand the stacked report icons and the curve representations, citing reasons such as familiarity with the representation, “the representation shows the number of students in each range,” “the representation provides a symbol per each 10 students,” and layout issues.

Sample explanations include “This is the most simplified version. Many teachers are afraid of graphs; this one is less threatening [stacked report icons]” (familiarity); “In my district, version A [box-and-whisker] because that is what we have been exposed to. Version B [stacked report icons] is pretty easy to understand too” (familiarity); “I think other teachers would be able to relate this display [curve] to the concept of a bell curve in their grading. The box-and-whisker

plot is definitely not going to be clear to other teachers” (familiarity); “The number of students in each quartile is given to you, so you don't have to do any calculations to figure them out [curve]” (shows number of students in each range); “I believe this one [curve] would be the easiest to understand for other students because it gives the numbers and the percentiles the clearest” (shows number of students in each range); “I think most teachers would prefer the bar graph [stacked report icons] because I think it is easier to determine scores” (shows symbol per each 10 students); “Most teachers look for the clearly defined and concise method of counting scores without doing too much calculating [stacked report icons]” (shows symbol per each 10 students); “This [stacked report icons] shows students and even if you do not understand percentiles you can see where your students fall and how many scored in each category (basic, proficient, etc)” (shows symbol per each 10 students); “I think teachers would like the visual representation [stacked report icons]” (layout issues); “Overall, I like version C [curve]. It is simple to understand, has color, and the shape of the curve helps to display the trend” (layout issues); and “[curve] It is the most visually clear” (layout issues).

Table 11

Explanation Categories and Counts for Question 3a, “Which Version Makes It Easier for You to Identify the Number of Students Whose Score Fell Within a Specific Range?”

Explanation category	Preferred: box-and- whisker	Preferred: stacked report icons	Preferred: curve	Total
Familiarity	0	0	1	1
Shows symbol per each 10 students/I can count	0	38	2	40
Shows number of students in each range (performance level/percentile)/no need to count	1	5	49	55
Layout issues: use of color, easier to read, user friendly, clearer, visually attractive, clear text explanations, and ordered	1	12	35	48
N/A, Other	2	4	0	6

Note. An explanation can belong to more than one category.

Finally, participants were asked to provide their suggestions on how these score reports could be made more effective. Table 13 shows the number of occurrences in each category. Suggestions provided include issues such as adding definitions, examples, explanations, and a summary; showing the number of students on each performance level; reducing the amount of text; and improving layout.

Table 12

Categories and Counts for Question 4a, “Which Version Do You Think Other Teachers Would Be Able to Understand the Best?”

Explanation category	Preferred: box-and- whisker	Preferred: stacked report icons	Preferred: curve	Total
Familiarity	1	3	20	24
Shows number of students in each range (performance level/percentile)/no need to count	0	6	15	21
Shows symbol per each 10 students/I can count	0	21	0	21
More information/more details	0	5	3	8
Less information/less details	0	0	1	1
Seems more real/seems more exact	0	2	0	2
Layout issues: use of color, easier to read, user friendly, clearer, visually attractive, clear text explanations, and ordered	3	24	30	57
N/A, Other	3	5	5	13

Note. An explanation can belong to more than one category.

Some of the suggestions include “Give a brief description how the mean is figured and why it is used” (add definitions); “Provide clear guidelines for their interpretation as not all teachers will be mathematically inclined” (add definitions); “For a non-math person, showing the number of students in each quartile may be helpful” (show number of students per quartile); “They would be more effective if the percentile lines were further apart and a different color than the median lines” (improve layout); and “The pictorial representations are easier to interpret” (OK as they are).

Table 13

Categories and Counts for Question 5, “How Could These Score Reports Be Made More Effective?”

Explanation category	Preferred: box-and- whisker	Preferred: stacked report icons	Preferred: curve	Total
Add examples	3	3	2	8
Add definitions, explanations for how to read the representation, provide a summary	10	6	9	25
Add a graph with all of the scores	1	1	3	5
Provide learning materials, information for formative purposes	3	0	1	4
Show number of students on each performance level	8	3	6	17
Reduce the amount of text	2	2	5	9
Improve layout: color use/graphics/text explanations	7	15	12	34
OK as they are	7	12	6	25
N/A, Other	8	8	5	21

Note. An explanation can belong to more than one category.

Summary and Discussion

Comprehension results and item difficulty analysis showed that the stacked report icons and the curve representations provide participants with additional support for answering questions about distributions. These representations either show the number of students in a particular range (e.g., between percentiles) or show icons that participants can count to determine the number of students that are in a range. These representations also show the shape of the distribution. The box-and-whisker representation that is used in many score reports does not provide this type of support. Enhancing this representation by adding the number of students in a range could help alleviate some of these issues.

Participants’ preference for a particular representation was influenced by several aspects including familiarity, more information being provided by the representation, and layout issues.

Participants who reported greater comfort with statistics tended to have higher comprehension scores.

Participants valued a particular representation based on their own knowledge level and on how useful the representation was at helping them answer the comprehension questions correctly. Both prior knowledge and the affordances of the graphical representation play a role in clearly communicating assessment results to particular audiences.

Future Work

Future work includes making changes to current score report representations of group performance and exploring comprehension and preference aspects of other external representations.

References

- Bagnall, L. (1994). Tabletop and Tabletop Jr.: Two tools for hands-on data exploration for kids. In C. Plaisant (Ed.), *Conference on Human Factors in Computing Systems, CHI 1994*, Boston, MA (pp. 63–64). doi: 10.1145/259963.260043
- Bakker, A., & Gravemeijer, K. P. E. (2005). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147–168). New York, NY: Kluwer Academic Publishers.
- Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Assessment issues of the 21st century* (pp. 43–61). New York, NY: Springer.
- Cobb, P. (1999). Individual and collective mathematical development: The case of statistical analysis. *Mathematical Thinking and Learning*, 1, 5–43.
- 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.
- Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report no. 430). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.
- Key Curriculum Press. (2011). *Fathom[®] dynamic data*. Retrieved from <http://www.keypress.com/x5656.xml>
- 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.
- Roberts, R., & Gierl, M. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice*, 29(3), 25–38.
- Spear, M. E. (1952). *Charting statistics*. New York, NY: McGraw-Hill.
- Steinke, T. (2005). TinkerPlots turns students into data analysts. *T H E Journal*, 32(9), 44.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Pearson.

- VanWinkle, W., Vezzu, M., & Zapata-Rivera, D. (2011, April). *Question-based reports for policymakers*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Vezzu, M., VanWinkle, W., & Zapata-Rivera, D. (2012). *Designing and evaluating an interactive score report for students* (Research Memorandum No. RM-12-01). Princeton, NJ: Educational Testing Service.
- Zapata-Rivera, D. (2011). Designing and evaluating score reports for particular audiences. In D. Zapata-Rivera & R. Zwick (Eds.), *Test score reporting: Perspectives from the ETS score reporting conference* (Research Report No. RR-11-45). Princeton, NJ: Educational Testing Service.
- Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2010, May). *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, Denver, CO.
- Zenisky, A., & Hambleton, R. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, 31(2), 21–26
- Zwick, R., Sklar, J., Wakefield, G., Hamilton, C., Norman, A., & Folsom, D. (2008). Instructional tools in educational measurement and statistics (ITEMS) for school personnel: Evaluation of three web-based training modules. *Educational Measurement: Issues and Practice*, 27, 14–27.

Notes

- ¹ More information on CBAL can be found at www.ets.org/research/topics/cbal.
- ² Even though participants may be more familiar with a vertical representation of the box-and-whisker plot, we chose to represent it horizontally for consistency purposes to facilitate comparison with the other two representations.
- ³ *P* value refers to the proportion of participants answering a question correctly. *P* values range from 0.00 to 1.00.

Appendix A

Comprehension Questionnaire

ETS Listening. Learning. Leading.®

Welcome to Understanding Score Reports

You will have up to 1 hour to complete this study.

There is no timer, but you will receive a warning 10 minutes before the hour is up if you have not finished by that time.

You must answer each question before proceeding to the next one. You will not have the opportunity to restart or to go back and change answers.

You may find these questions to be difficult, since we have intentionally provided only minimal information. We are not expecting participants to answer all the questions correctly, but please do your best.

Next

Copyright © 2011 by Educational Testing Service. All Rights Reserved.

Understanding Score Reports

Instructions

You may find some of these questions to be difficult to answer based on the information provided in the figure. You will get to explore other representations at the end of the study.

For questions 1-14, choose the best answer to the question by looking at the score report display on the right.

Question 1

1. Which of the following statements is true?

- The same amount of students scored between 180 and 280 as above 280.
- The same amount of students scored between 180 and 280 as below 180.
- More students scored between 180 and 280 than below 180.
- Less students scored between 180 and 280 than below 180.

CBAL Group Performance

MATH Score Reports

Teacher: Ms. Brown
Grade: 8
Test Date: 09/30/11

How did the 8th-graders in the School do on the Math Test?

A total of 240 students completed the test.

Lowest: 50
25th Percentile*: 180
Median: 230
75th Percentile*: 280
Highest: 390

*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.

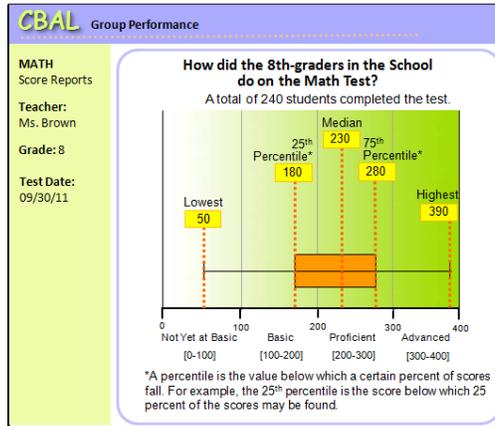


Understanding Score Reports

Question 2

2. Which of the following statements is true?

- The scores are equally distributed across performance levels.
- More than 25% of the scores are at the Advanced level.
- 25% of the scores are at the Not Yet at Basic level.
- More than 25% of the scores fall in the Proficient level.

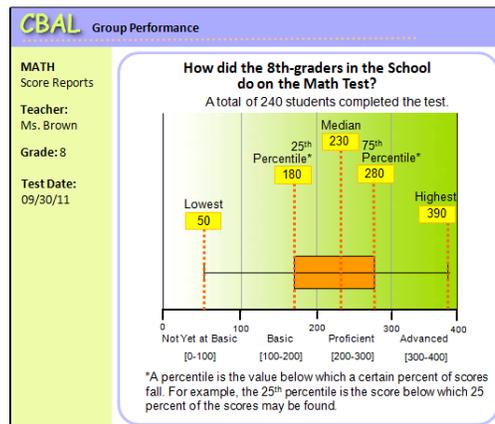


Understanding Score Reports

Question 3

3. How many students scored between 50 and 180?

- 25
- 60
- 130
- 180



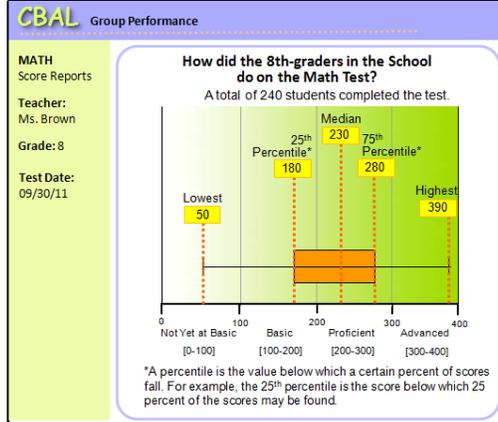


Understanding Score Reports

Question 4

4. How many students scored between the 25th percentile and the 75th percentile?

- 50
- 100
- 120
- 230

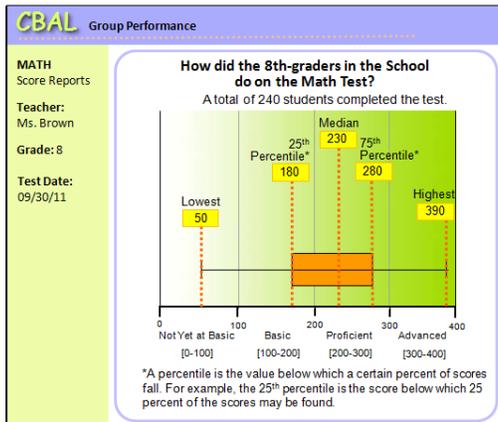


Understanding Score Reports

Question 5

5. How many students scored below 280?

- 60
- 120
- 180
- 280



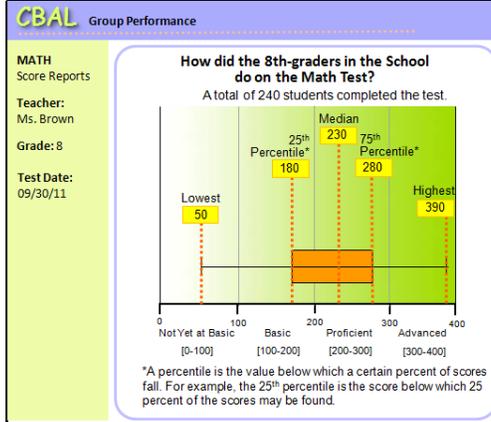


Understanding Score Reports

Question 6

6. How many students scored between 280 and 390?

- 50
- 60
- 120
- 180

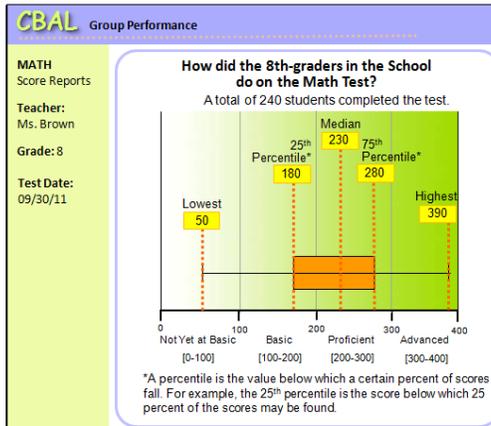


Understanding Score Reports

Question 7

7. How many students scored between the 25th percentile and the highest score?

- 60
- 120
- 180
- 240

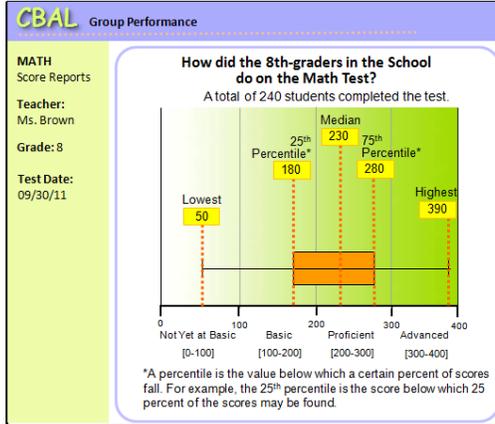




Understanding Score Reports

Question 8

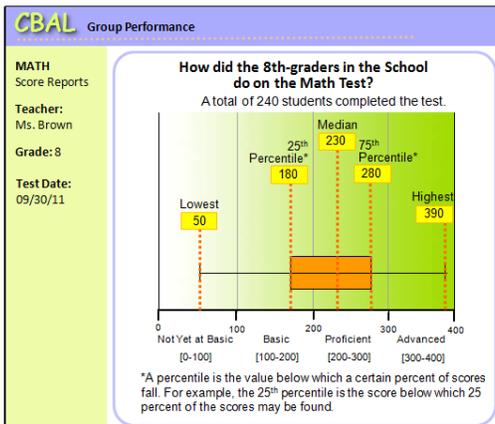
8. Which of the following statements is true?
- 60 of the students' scores fall below the median score.
 - 144 of the students' scores fall between the median and highest score.
 - 120 of the students' scores fall at or below the median score.
 - 180 of the students' scores fall above the median score.



Understanding Score Reports

Question 9

9. What is the score range where 50% of the scores fall according to the graph?
- 50-180
 - 180-280
 - 180-390
 - 280-390



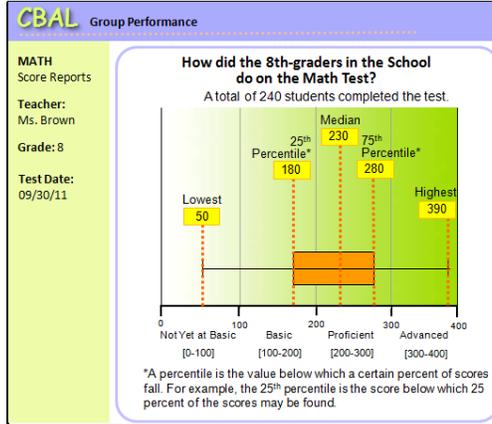


Understanding Score Reports

Question 10

10. Assume a student in the school, Pat, scored 220 on the test. What can we conclude about his score in comparison to other 8th graders who took the test?

- 50% or more of the students scored lower than Pat.
- 25% of the students scored above Pat.
- 25% of the students scored below Pat.
- 50% or more of the students scored higher than Pat.

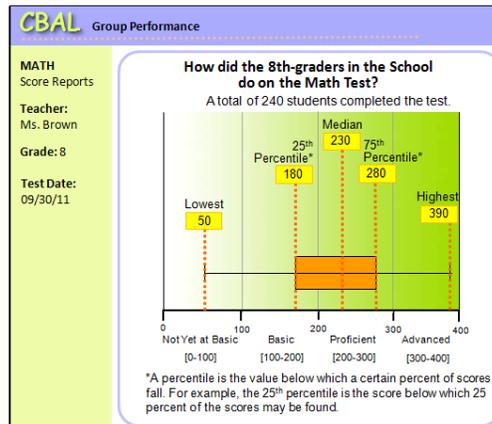


Understanding Score Reports

Question 11

11. What is the score range where 50% of the scores fall according to the graph?

- 50-280
- 50-180
- 230-390
- 280-390



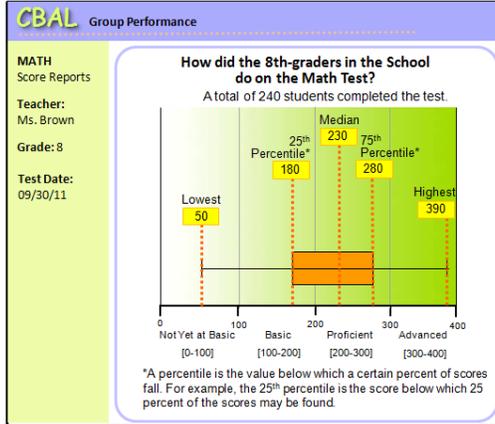


Understanding Score Reports

Question 12

12. Assume a student in the school, Pat, scored 220 on the test. Based on the graph, we can say that Pat's score falls

- between the median and the highest score.
- between the 75th percentile and the highest score.
- between the 25th and the 75th percentile.
- between the lowest and the 25th percentile.

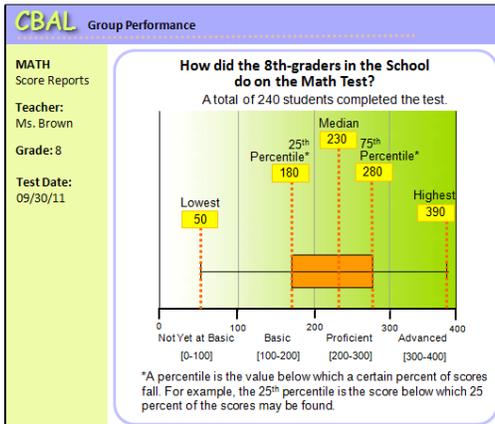


Understanding Score Reports

Question 13

13. Assume another student in the school, Juanita, scored 150 on the test. Which of these statements is true based on the graph?

- Juanita's score is higher than 25% of the 8th graders' scores in the school.
- 50% or less of the students scored higher than Juanita.
- Juanita's score falls below the 25th percentile.
- Juanita scored higher than 25% of the students on the test.



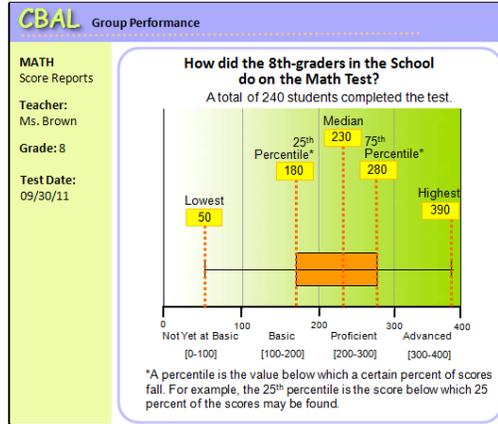


Understanding Score Reports

Question 14

14. Assume another student in the school, Terrell, scored 330 on the test. Based on the graph, which statement is correct?

- Terrell's score is between the median and 75th percentile.
- Terrell's score is below 75% of the scores of all 8th graders.
- Terrell's score is the highest on the test.
- Terrell's score is higher than 75% of the scores of all 8th graders.



Appendix B Preference Questionnaire



In the next set of questions, you will see three different versions of a score report and make judgments about them.

Continue

Copyright © 2011 by Educational Testing Service. All Rights Reserved.

Understanding Score Reports

1a. Here are three different versions of a score report showing how 8th graders performed on a math test. **Which one do you prefer?**

Please look at all the score report displays before answering the question. Click on a smaller version to enlarge it.

Please select your preference: A B C

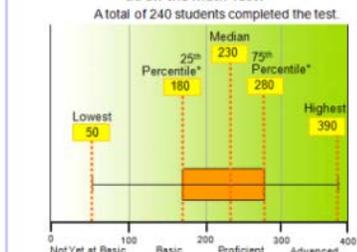
CBAL Group Performance

MATH Score Reports

Teacher: Ms. Brown
Grade: 8
Test Date: 09/30/11

How did the 8th-graders in the School do on the Math Test?

A total of 240 students completed the test.

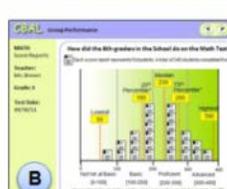


*A percentile is the value below which a certain percent of scores fall. For example, the 25th percentile is the score below which 25 percent of the scores may be found.

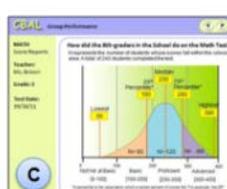
Next Question



A



B



C

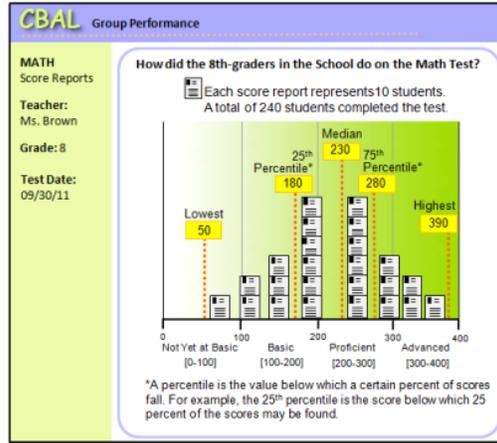
Understanding Score Reports

1a. Here are three different versions of a score report showing how 8th graders performed on a math test. **Which one do you prefer?**

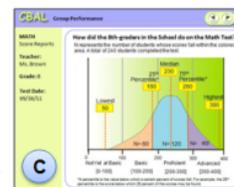
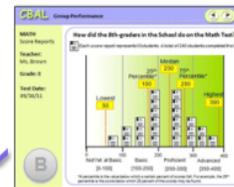
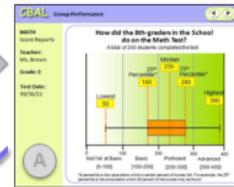
Please look at all the score report displays before answering the question. Click on a smaller version to enlarge it.

Please select your preference: A B C

Next Question



1b. Explain why you prefer this version.



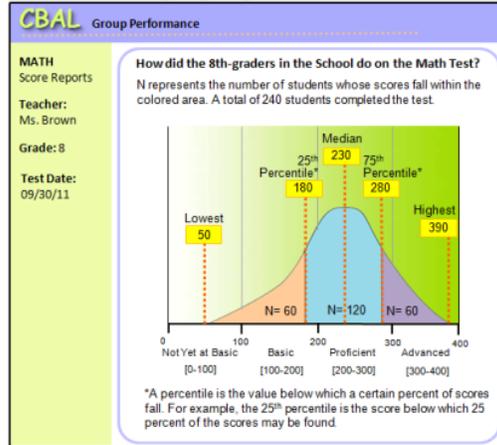
Understanding Score Reports

1a. Here are three different versions of a score report showing how 8th graders performed on a math test. **Which one do you prefer?**

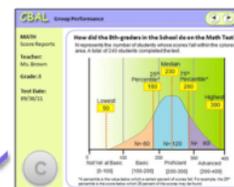
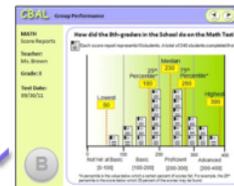
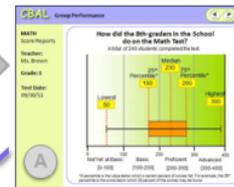
Please look at all the score report displays before answering the question. Click on a smaller version to enlarge it.

Please select your preference: A B C

Next Question



1b. Explain why you prefer this version.



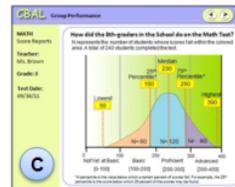
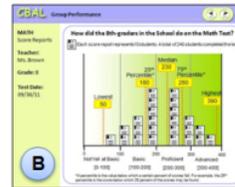
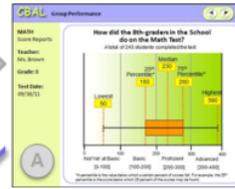
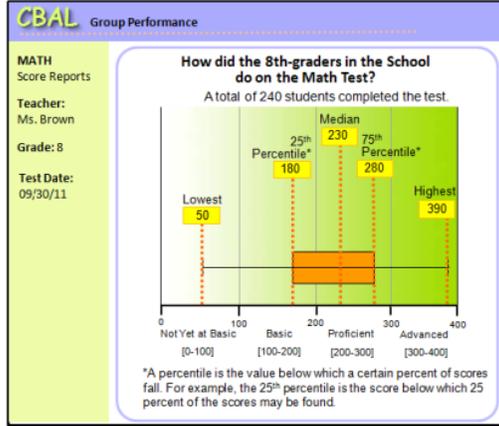
Understanding Score Reports

2a. Which of these three versions helps you understand the distribution better?

Please look at all the score report displays before answering the question. Click on a smaller version to enlarge it.

Next Question

Please select your preference: A B C



2b. Explain why you prefer this version.

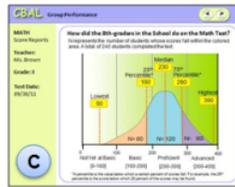
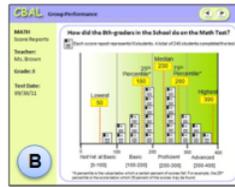
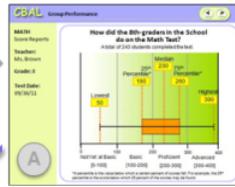
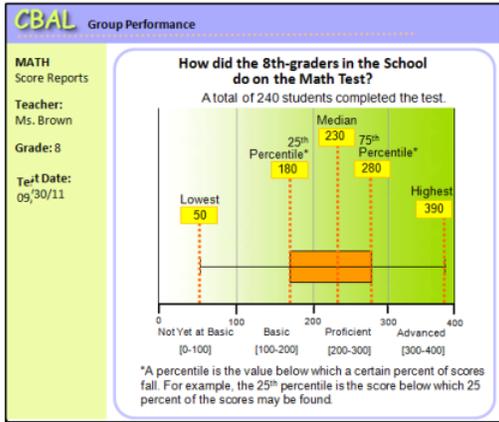
Understanding Score Reports

3a. Which version makes it easier for you to identify the number of students whose score fell within a specific range (e.g., 180 students scored between the lowest score and the 75th percentile)?

Please look at all the score report displays before answering the question. Click on a smaller version to enlarge it.

Next Question

Please select your preference: A B C



3b. Explain why this version makes it easier.

Understanding Score Reports

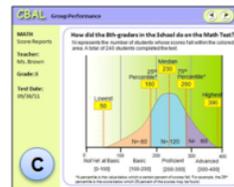
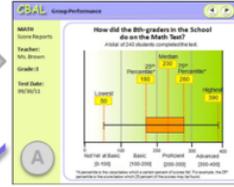
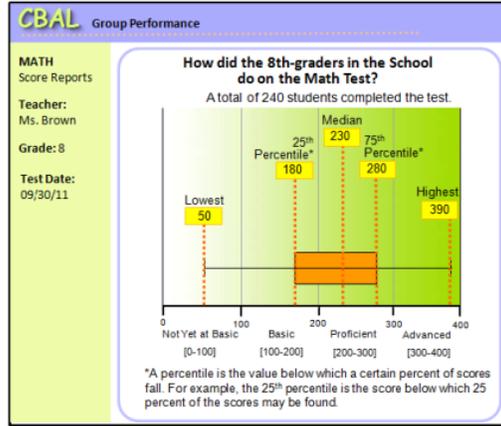
4a. Which version do you think other teachers would be able to understand the best?

Please look at all the score report displays before answering the question. Click on a smaller version to enlarge it.

Next Question

Please select your preference: A B C

A



4b. Explain why.

Understanding Score Reports

5. How could these score reports be made more effective?

Next Question