RESEARCH REPORT 2013-3

CollegeBoard

Rating Quality Studies Using Rasch Measurement Theory

By George Engelhard, Jr. and Stefanie A. Wind

George Engelhard Jr. is a professor of educational measurement and policy in the Division of Educational Studies at Emory University in Atlanta, Ga.

Stefanie A. Wind is a doctoral student in educational measurement and policy in the Division of Educational Studies at Emory University in Atlanta, Ga.

Acknowledgments

Paper presented at the annual meeting of the American Educational Research Association in Vancouver (April 2012). The College Board provided support for this research. Researchers supported by the College Board are encouraged to freely express their professional judgment. Therefore, the points of view or opinions stated in College Board–supported research do not necessarily represent official College Board position or policy.

About the College Board

The College Board is a mission-driven not-for-profit organization that connects students to college success and opportunity. Founded in 1900, the College Board was created to expand access to higher education. Today, the membership association is made up of over 6,000 of the world's leading educational institutions and is dedicated to promoting excellence and equity in education. Each year, the College Board helps more than seven million students prepare for a successful transition to college through programs and services in college readiness and college success — including the SAT® and the Advanced Placement Program®. The organization also serves the education community through research and advocacy on behalf of students, educators and schools. For further information, visit www.collegeboard.org.

For more information on College Board research and data, visit research.collegeboard.org.

Contents

Tables

Figures

Executive Summary

The major purpose of this study is to examine the quality of ratings assigned to CR (constructed-response) questions in large-scale assessments from the perspective of Rasch Measurement Theory. Rasch Measurement Theory provides a framework for the examination of rating scale category structure that can yield useful information for interpreting the meaning of ratings assigned in large-scale performance assessment contexts. This study uses data collected as a part of the reader reliability studies (Miao & Odumade, 2011) based on the 2010 administration of the AP® Statistics Exam (N = 238 students, N = 156 raters). The following research questions are addressed: (1) Do the raters on the AP Statistics Exam vary in severity? (2) Do the CR questions on the AP Statistics Exam vary in difficulty? (3) Is the structure of the rating scale comparable across the CR questions?

The AP Statistics Exam consists of a combination of MC (multiple-choice) items (40 items) and six CR questions related to statistical content domains. The College Board currently examines rating quality for CR questions with Generalizability Theory. This study augments and extends current research on reader reliability with indicators of rating quality based on Rasch Measurement Theory.

Rating Quality Studies Using Rasch Measurement Theory

Rater-mediated assessments play an important role in the College Board's Advanced Placement Program® (AP). Most of the AP Exams consist of a mixture of various item types. For example, the AP Exam in statistics consists of MC items and CR questions. There are a variety of procedures for evaluating the psychometric quality of mixed-mode assessments like the AP Exams. These procedures are based on rater agreement, Generalizability Theory, and various item response theory models. Currently, the psychometric quality of rater judgments for AP Exams is examined using rater agreement indices and Generalizability Theory (Miao & Odumade, 2011).

The purpose of this study is to examine the psychometric quality of the ratings assigned to CR questions on the AP Statistics Exam using indices of rating quality based on Rasch Measurement Theory. Guidelines recommended by Linacre (1999, 2002) and Engelhard (2002) for the interpretation of Rasch-based indices of rating quality are used to explore the psychometric quality of the rating scales for each of the six CR questions. In addition, graphical displays from the Facets computer program (Linacre, 2010) are compared across the six CR questions in order to compare the structure of the AP Statistics rating scale across these items. The following research questions guide this study:

1. Do the raters on the AP Statistics Exam vary in severity?
2. Do the CR questions on the AP Statistics Exam vary in difficulty?
3. Is the structure of the rating scale comparable across the CR questions?

The overall goal of this study is to explore the value of the additional quality-control information provided by the Many-Facet Rasch (MFR) Model (Linacre, 1989) in order to improve the rater-mediated aspects of the College Board Advanced Placement® Exam system.

Theoretical Framework: Rasch Measurement Theory

Item Response Theory is the general theoretical framework for this study. Specifically, principles from Rasch Measurement Theory and the MFR model guide the analyses and interpretation of the data in this study. The Rasch model and its extensions conceptualize measurement in terms of a latent variable represented as a single line that can be used to explain differences in person achievement and item difficulty. The family of Rasch models (Rasch, 1960/1980) is a set of probabilistic Item Response Theory (IRT) models that meet the requirements for invariant measurement (Engelhard, 2008). When acceptable model-data fit is observed, the requirements for invariance have been met. In other words, a match between observations in data and expectations by the Rasch model indicates that the construct of interest is being measured without interference by construct-irrelevant factors, such as individual rater characteristics, or dependence on construct-irrelevant characteristics of an assessment situation, such as persons or items.

In contrast to Classical Test Theory, in which raw or total scores are used to describe student achievement, Rasch Measurement Theory uses a logistic transformation to convert raw-score observations to measures on a log-odds scale. When test score data are modeled using this transformation, measures on the log-odds scale can be viewed as the dependent variable with various facets, such as student achievement, rater severity, domain difficulty, and task difficulty, conceptualized as independent variables that influence these log-odds. Based on the difference between item calibrations and person locations on the logit scale, items can be judged in terms of their usefulness for providing information about specific persons at varying levels on the latent variable. The logit-scale location of elements within each facet can be compared visually using the graphical display provided in most Rasch measurement software, including Facets (Linacre, 2010), which is the program used in this study. The Facets program produces a *variable map* in the output for Rasch analyses, which displays location estimates for each facet on a common vertical ruler. Examples of variable maps from Facets are provided in Figures 1 and 2; these displays are described in the results section.

Various models based on Rasch Measurement Theory can be used to describe relationships among facets of an assessment situation, including students and items. In this study, the MFR model (Linacre, 1989) is used to estimate the location of facets in a performance-assessment context. The MFR model is an extension of Rasch Measurement Theory models used to depict the relationship between facets of an assessment situation as well as the probability (described in terms of the log of the odds, or logit) for observing specific outcomes within assessment situations involving multiple facets. Variations on the MFR model allow the structure of a rating scale to be fixed across items (the rating scale MFR model), or to vary for each item (the partial-credit MFR model). In the context of a rater-mediated assessment, the rating scale MFR model can be expressed as:

$$\ln\left[\frac{P_{nijkx}}{P_{nijkx-1}}\right] = \theta_n - \lambda_i - \delta_j - \Delta_k - \tau_x \qquad (1)$$

where

P_{nijx} = probability of student n receiving a rating of x by rater i on question j,

P_{nijx-1} = probability of student n being rated $x-1$ by rater i on question j,

θ_n = statistics achievement of student n,

λ_i = severity of rater i,

δ_j = difficulty of MC item j,

Δ_k = difficulty of CR question k, and

τ_x = difficulty of category x relative to category $x-1$.

The partial-credit MFR model can be expressed as:

$$\ln\left[\frac{P_{nijkx}}{P_{nijkx-1}}\right] = \theta_n - \lambda_i - \delta_j - \Delta_k - \tau_{kx} \qquad (2)$$

where

P_{nijx} = probability of student n receiving a rating of x by rater i on question j,

P_{nijx-1} = probability of student n being rated $x-1$ by rater i on question j,

θ_n = statistics achievement of student n,

λ_i = severity of rater i,

δ_j = difficulty of MC item j,

Δ_k = difficulty of CR question k, and

τ_{kx} = difficulty of category x relative to category $x-1$ for CR question k.

The major difference between the rating-scale and partial-credit MFR models is related to the category coefficient locations, which are indicated in the model with tau (τ). The tau is not a facet, but represents the difference in difficulty (or location on the latent variable) between adjacent categories in a rating scale. When data are modeled using the rating scale MFR model, category coefficient locations are fixed across items, indicated by the τ_x term. As a result, the distance on the latent variable between each pair of rating scale categories does not vary across tasks. Panel A of Table 1 provides a visual representation of this concept. Each of the four rating scale category coefficients is separated by equally spaced intervals on the latent variable. When this model is used, the distance between two adjacent categories is considered equivalent to the distance between any other two adjacent categories.

In contrast, the partial-credit formulation of the MFR model allows category coefficient locations to vary across items, indicated by the τ_{jx} term in the model. The partial-credit model is a useful diagnostic tool for comparing rating scale category use across a set of items scored by raters; it is essentially a test of the hypothesis of equidistant categories across items. As illustrated in Panel B of Table 1, this assumption may not be reflected in the observed use of rating scale categories. The partial-credit MFR model allows this hypothesis to be empirically investigated.

Rating Scale Category Guidelines

Unlike MC items, rater-mediated assessments require raters to identify characteristics of a response that suggest student locations on a latent variable. According to Andrich, de Jong, and Sheridan (1997), rating scales “partition a latent unidimensional continuum into adjacent intervals” (p. 59) that provide raters with a format for describing their judgments according to criteria described in a rubric. Typical response features and examples are often provided during rater training to help scorers become familiar with qualitative differences in performance associated with each level on a rating scale.

Linacre (1999, 2002) describes a set of guidelines for examining the quality of rating scales using Rasch Measurement Theory. Based on Linacre’s work, a summary of these guidelines is given in Table 2. Adherence to the guidelines by a particular data set suggests that rating scale categories are functioning as intended based on the model, and that they can be used to describe student locations on a latent variable (Linacre, 1999). In the following sections, each of these seven guidelines for evaluating rating scale quality is described.

1. **Directionality.** The first guideline requires directional orientation of sequential rating scale categories with the latent variable. In other words, when the directionality of a rating scale is oriented with the latent variable, high ratings and high values of summed ratings reflect high locations on the latent variable. A bivariate plot of observed and expected measures provides information about directionality in a rating scale that can be used to support inferences about the progression of difficulty implied by ordered categories. Linacre (2002) describes directionality in terms of a match between observed ratings and the corresponding model predicted (expected ratings), which he calls the coherence of category usage.

2. **Monotonicity.** Along the same lines, Guideline 2 requires monotonic progression of rating scale categories. Average person locations across all observations in each category can be used as evidence of monotonicity. When this guideline is met, increasing rating scale categories will correspond to increasing average person measures on the latent variable within the categories.

3. **Category Usage.** The frequency and distribution of observed ratings across rating scale categories is used as evidence of adherence to Guidelines 3 and 4. When the frequency of observations across rating scale categories is not equal, the categories may not indicate substantive differences in the meaning of ratings (Linacre, 2002). Because the estimation of category coefficient locations from Rasch Measurement Theory depends on frequencies of observations within rating scale categories, Linacre (2002) has suggested as a rule of thumb that categories with fewer than 10 observations limit the precision and stability of these estimates. Unobserved categories present significant challenges to the interpretation of rating scales. Categories with no observations must be distinguished as either structural or incidental zeroes. A structural zero occurs when category requirements are impossible to fulfill, and an incidental zero occurs when an unobserved category is the consequence of a particular sample. Linacre (2002) describes strategies for addressing issues related to unobserved categories.

4. **Distribution of Ratings.** Guideline 4 is directly related to the requirements specified for category usage by Guideline 3 and focuses on the percentage of observations within rating scale categories for a given task. Adherence to this guideline is observed when ratings conform to a regular distribution, such as uniform, normal, unimodal, or bimodal distributions. The presence of skew or modality in graphical displays of rating distributions across categories can be used to quickly identify violation of this guideline.

5. **Rating Scale Fit.** Guideline 5 is related to unexpected use of rating scale categories. Indices of model-data fit allow for a comparison of the randomness observed in data with the randomness expected by the model. When data fit the Rasch model, a reasonably uniform level of randomness will be observed. Mean square error (*MSE*) statistics for residual analyses provide information about model-data fit. Outfit *MSE* statistics are particularly useful in this context because of their sensitivity to outliers. With an expected value around 1.00, low Outfit *MSE* statistics suggest "muted" rating patterns, or rating data that are more uniform than predicted. In contrast, high values suggest excessive randomness or overly "noisy" rating patterns and indicate the use of categories in unexpected contexts.

6. **Category Coefficient Order.** In order to meet the requirements of Guideline 6, category coefficient locations must reflect the intended order of categories in terms of progression or development along the latent variable (Andrich, 1978a; 1978b). In addition, the "inferential interpretability" and the ability for a rating scale to produce invariant measures depends on a monotonic sequence of category coefficient locations (Linacre, 2002). Comparison of these locations on a variable map can be used to identify the match between observed and intended category ordering (Engelhard, 2002).

7. **Category Coefficient Locations.** Guideline 7 is related to the precision of rating scale categories for the description of person performance at different locations on the latent variable. Huynh (1994) and Linacre (1997; 2002) describe a range of threshold location differences between 1.40 and 5.00 logits ($1.40 < \tau_k - \tau_{k-1} < 5.00$ logits) as evidence that rating scale categories are distinctive. When categories are distinctive, each describes a unique range of person locations on the latent variable. Peaked information functions that appear as a "range of hills" provide evidence for distinctive category coefficient locations across rating scale categories (Linacre, 2002). Information functions for rating scale categories are further described in the next section.

Graphical Displays for Examining Rating Quality

In addition to the quantitative indices of rating quality, there are seven graphical displays that can be examined in order to evaluate rating scale categories. Specifically, these are: (1) Category Probability Functions, (2) Expected Score ICC, (3) Expected Score with Empirical ICC, (4) Cumulative Probabilities, (5) Conditional Probabilities, (6) Item Information, and (7) Category Information. In this section, the interpretation of each display is described. Detailed interpretations of these graphical displays related to the current data analyses appear in the results section. Displays for the ratings in the AP Statistics Exam were created using the Facets program (Linacre, 2010).

1. **Category Probability Functions.** Category probability functions are a visual representation of the probabilistic relationship between category difficulty and student location on the latent variable. Each curve represents an individual rating scale category, and the curves always appear in ascending order so that the curve representing the lowest category is farthest to the left and the curve for the highest category is farthest to the right. When rating scale categories function as intended, increasing student location on the latent variable is associated with an increasing probability for observed ratings in higher categories.

   Category probability functions are particularly useful for comparing the structure of rating scales when a partial-credit model is used. Because the partial-credit MFR model does not require fixed category widths, the location of category curve peaks may vary across items. Although the formulation of the category probability curves is such that they always appear in ascending order from left to right, the crossover points between adjacent curves can be disordered if the scale is not functioning as intended. In this situation, categories that are never the most probable at any point along the *x*-axis (i.e., are non-modal) do not have distinct peaks.

2. **Expected Score ICC.** The category probability curves for a rating scale item combine to produce a single Item Characteristic Curve (ICC), which is a Rasch model logistic ogive. Like the category probability curves, the ICC graphically represents probabilistic relationship between student location on the latent variable and the probability for observed ratings in given categories. This display is particularly useful for predicting student measures as they relate to individual rating scale categories. Dashed lines displayed on the plot relate the Rasch half-point thresholds for each category (*y*-axis) to an interval on the *x*-axis for which a range of measures round to a corresponding rating scale category.

3. **Expected-Score-with-Empirical ICC.** The expected-score-with-empirical ICC is interpreted in a similar fashion to the expected ICC described above. The utility of this display is its joint demonstration of empirical (observed) ratings and model expectations. The *X*s identify the observed average rating on the *y*-axis for an interval of student measures on the latent variable (*x*-axis). Confidence bands are drawn around the curve and represent upper and lower bounds of a 95% confidence interval. Observations that fall outside these bands indicate misfit, or unexplained variance.

4. **Cumulative Probabilities.** The cumulative probability curves display model-based cumulative probabilities for rating scale categories in terms of intervals along the latent variable (x-axis). With the lowest and highest categories represented by the farthest-left and farthest-right curves, respectively, each curve displays the probability (y-axis) for students at given points on the latent variable (x-axis) to receive a rating in a given category or the one just below it. For example, the third category curve is the model-based sum of the probabilities for an observed rating of the first three categories (probability for category 1+2+3). The distance between each curve corresponds to an interval of student locations on the x-axis and can vary across items in the partial-credit model. These curves are always in ascending order, and a dashed horizontal line is used to display the location of the Rasch-Thurstone threshold for each rating scale category. These thresholds are the point of intersection between a location on the latent variable (*x*-axis) and the 0.50 probability point (on the *y*-axis). For each category, the Rasch-Thurstone threshold is the point at which the probability for an observed rating in category k is equivalent to the probability for a rating in category $k-1$.

5. **Conditional Probabilities.** The conditional probability curves follow dichotomous logistic ogives and display the model-based relationship between probabilities for observed ratings in pairs of adjacent categories. Each curve represents two categories, such that the curve farthest to the left models the probability for a rating of the lowest and next-lowest categories (e.g., categories 0+1) for students with given locations on the latent variable (*x*-axis). In the Facets program (Linacre, 2010) output, a dashed horizontal line intersects each curve at the 0.50 probability point to indicate the location of the Rasch-Andrich threshold for each pair of categories. This is the point on the latent variable at which a category is most probable. These curves can be disordered.

6. **Item Information.** The item information curve displays the amount of model-based Fisher statistical information provided by an item at different locations on the latent variable (Fisher, 1958). Item information is related to the match between person location and item difficulty, and well-targeted items provide more information than items that are far from person locations. Along the same lines, item information is directly related to the precision of measurement, such that measures with small standard errors contribute more information than measures with large standard errors. The plot of item information for rating scale items can be used to identify locations along the latent variable at which the information is most useful for providing statistical information, identified by high values on the y-axis.

7. **Category Information.** The category information curves can be interpreted similarly to the item information curve described above. In this display, the amount of information provided across student locations on the latent variable is plotted separately for each rating scale category. For example, when a rating scale is functioning as intended, lower categories will provide more information for students with low measures on the latent variable than for students with high measures.

Methods

Instrument

Data used in this study were collected during the 2009 administration of the Statistics Exam. This exam is administered at the conclusion of the AP Statistics course. The AP Program, created and maintained by the College Board, provides high school students with college-level course work in a variety of subjects, with opportunities to earn college credit, placement, or both, based on examination performance. Each AP Exam is developed by subject-matter experts from secondary and postsecondary institutions and organizations, and they are designed to reflect the content and academic rigor of college-level courses in specified subject areas. AP Exams are offered in a variety of subjects that match a wide range of academic interests.

Data used in this study were collected during the 2009 administration of the Statistics Exam. This exam is administered at the conclusion of the AP Statistics course. The AP Program, created and maintained by the College Board, provides high school students with college-level course work in a variety of subjects, with opportunities to earn college credit, placement, or both, based on examination performance. Each AP Exam is developed by subject-matter experts from secondary and postsecondary institutions and organizations, and they are designed to reflect the content and academic rigor of college-level courses in specified subject areas. AP Exams are offered in a variety of subjects that match a wide range of academic interests.

AP Statistics Course Background. The AP Statistics course and exam were created by statistics and mathematics educators to match the typical content included in college-level introductory statistics courses. The course is designed to introduce students to principles and concepts that allow them to collect, analyze, and draw conclusions from data based on four conceptual themes:

1. Exploring Data: Describing patterns and departures from patterns

2. Sampling and Experimentation: Planning and conducting a study

3. Probability and Simulation (two questions): Exploring random phenomena using probability and simulation

4. Statistical Inference (two questions): Estimating population parameters and testing hypotheses (*AP Statistics Course Description*, 2010)

Although there is some variation, most AP Statistics courses are offered as a one-year, or two-semester, course for secondary school students who have successfully completed prerequisite requirements in mathematics. The AP Exam is offered at the end of the course, and scores are used to qualify students for college placement, credit, or both. Detailed information about the scope and sequence of content provided in a typical AP Statistics course is given in the *AP Statistics Course Description* (College Board, 2010).

AP Statistics Exam. The AP Statistics Exam consists of two sections: 40 MC items and six CR questions. These two sections are combined to produce a composite score on a 5-point scale, which is used to reflect a student's qualification for course credit or placement at the college level. A score of 5 indicates the highest level of qualification and is assumed to reflect a score of A in a corresponding college course. A score of 1 indicates the lowest level of qualification.

The AP Statistics Exam is designed to measure a student's mastery of concepts and techniques related to the subject matter from the AP Statistics course. The MC and CR portions of the exam are weighted equally, and students are given 90 minutes to complete each section. MC items are scored electronically and require students to select a single answer from five possible choices. In contrast, the CR section of the AP Statistics Exam is scored by human raters and requires students to answer five questions and complete an investigative task (*AP Statistics Course Description,* College Board, 2010). These six CR questions are designed to assess a student's ability to:

> "Relate two or more different content areas (i.e., exploratory data analysis, experimental design and sampling, probability, and statistical inference) as they formulate a complete response or solution to a statistics or probability problem" (p. 27)

and

> "Demonstrate their mastery of statistics in a response format that permits the students to determine *how* they will organize and present each response" (p. 27).

Raters who score CR questions on this exam are college faculty or highly qualified AP teachers hired by the College Board who are trained to score items according to specifications set forth in a rubric. CR questions are scored on a scale of 0–4, with a score of 4 reflecting a complete response and the highest possible score, and a score of 0 reflecting the lowest possible score (see Appendix for score point descriptions). This study focuses on the ratings assigned to CR questions. The six CR questions in the 2010 administration of the AP Statistics Exam were based on six questions (e.g., domains) associated with course content. The CR item topics are as follows:

1. Sampling and Experimentation

2. Probability and Simulation

3. Statistical Inference

4. Probability and Simulation

5. Statistical Inference
6. Exploring Data

Scoring the CR Questions. During rater training for scoring the AP Statistics Exam, raters are trained to use item-specific rubrics to assign scores to student responses to each of the six CR questions for a given administration. Specifically, raters assign scores to each response using a 0–4 scale using evidence of statistical knowledge and communication skills in student responses. Table A1 provides descriptors for each of the score levels related to the CR questions.

The scoring guidelines for the AP Statistics Exam are designed to reflect the complex nature of the CR questions (College Board, 2010). As a result, the scoring guidelines for this exam describe each CR question in two parts. Using examples and guidelines provided in the scoring manual, raters are instructed to evaluate student responses in both parts and determine if a student's response is Essentially Correct, Partially Correct, or Incorrect. Specific criteria for classifying performance are provided for each CR question, with item-specific examples and descriptions. Scores are determined based on performance in both parts. For example, both parts' *Essentially Correct* results in the highest score (4 points), and both parts' *Incorrect* results in the lowest score (0 points). Table A2 summarizes these scoring guidelines as they relate to the score categories for the AP Statistics Exam.

According to Engelhard (2009), rater-mediated assessments "do not provide direct information regarding student achievement because the student's responses must be mediated and interpreted through raters to obtain judgments about student achievement" (p. 261). Because CR questions do not have a single "correct" answer and therefore cannot be scored automatically, the role of human judgment must be considered in the interpretation of scores.

Participants

The data used in this study are a sample from the population of students and raters who participated in the 2009 administration of the AP Statistics Exam. The sample includes MC and CR item scores from 238 students, whose responses to the CR questions were scored by 156 raters. Demographic information for the student sample is provided in Table 3. In the reader reliability studies, conducted by Educational Testing Services (ETS) for the College Board, there are two individual raters scoring each CR item.

Procedures

In order to answer the research questions, models based on Rasch Measurement Theory are used to examine student achievement, rater severity, item and question difficulty, and rating scale structure using data from the AP Statistics Exam. Analyses are implemented using the Facets computer program (Linacre, 2010). Specifically, two models are used with these data. Model I is a rating scale MFR model and was presented in Equation 1; Model II is a partial-credit MFR model and was presented in Equation 2. Model I is used to explore the first research question, and analyses related to this model will include an examination of student, item, mode, and rater calibrations. A variable map is used to compare the location of each facet on a common scale.

Model II is related to the second research question and is used to examine the structure of the rating scale across the six CR tasks in the AP Statistics Exam. In order to examine the structure of the rating scales associated with each task in Model II, quantitative indices and graphical displays are examined and compared in terms of the rating scale category guidelines described earlier. A variable map that displays the rating scale category locations is also examined for this model.

Results

Model I

Variable Map. The variable map shown in Figure 1 presents a graphical display of the spread of student measures (statistics achievement), rater measures (severity), item and question calibrations (difficulty), and the location of the thresholds for the rating scale categories, all on the same logit scale. The first column shows the logit scale. The second column presents the student measures of writing achievement from the AP Statistics Exam. Higher scoring students appear at the top of the column, and lower scoring students appear near the bottom. Each asterisk represents two students, and a period represents one student. The student achievement measures in this sample range from –2.58 logits to 3.75 logits (M = 0.25, SD = 1.00, N = 239). Question calibrations are shown in the third column. The MC items are anchored at 0.00 logits, and the six other questions range from 0.38 logits for Question 6, which is the easiest, to 1.78 logits for Question 2, which is the hardest. Raters are shown in the fourth column and range in severity measures from −1.18 logits to 2.18 logits (M = 0.00, SD = 0.50, N = 156). Each asterisk represents one rater, with severe raters located near the top of the scale, and lenient raters located at the bottom. The fifth column shows the item difficulty calibrations on the logit scale, with item difficulty ranging from 0.00 logits to 1.78 logits (M = 0.00, SD = 1.29, N = 39). More difficult items are located near the top of the column, and easier items are located closer to the bottom. Finally, the rating scale structure is shown in the farthest-right column.

Summary Statistics. Table 4 provides summary statistics from Facets (Linacre, 2010) analyses for the students, MC items, CR questions, and raters examined in this study. The overall differences between students, items, CR questions, and raters are significant ($p < 0.05$), with high reliabilities of separation, although the reliability for raters is notably lower than for the other facets ($REL_{Students}$ = 0.92; REL_{Items}= 0.98; $REL_{Questions}$ = 0.99; REL_{Raters}=0.57). The reliability of separation statistics for persons from Facets (Linacre, 2010) is comparable to Cronbach's coefficient alpha. For other facets, the reliability of separation statistics describes the spread or differences between elements within a facet. The significant separation statistics indicate a spread of the elements within each of the facets across the latent variable (statistics achievement). Good fit to the model is evident for each of these facets, with mean Infit and Outfit *MSE* statistics near their expected values of 1.00. The fit statistics for the Rater facet have standard deviations about twice the size of those for the other facets examined using Model I (SD = 0.42 for Infit *MSE* and SD = 0.40 for Outfit *MSE*). Although the average values of the *MSE* statistics are close to their expected value, large standard deviations suggest that the values of fit statistics for several individual raters may be extreme when compared to the rest of the group. The calibration of individual raters under Model I is provided in Table 6; examination of this table reveals several raters with extremely low fit statistics (e.g., Rater 71533: Infit *MSE* = 0.15, Outfit *MSE* = 0.17), and several raters with extremely high values of fit statistics (e.g., Rater 17708: Infit *MSE* = 1.72, Outfit *MSE* = 2.08). Fit statistics below and above the expected value of 1.00 suggest that raters may be demonstrating rating patterns that are "muted" or more varied than expected, respectively.

Calibration of CR Items. Table 5 presents the calibration of the CR (CR) Items facet for the AP Statistics Exam data. For this analysis, the MC items are anchored at 0.00 in order to provide a frame of reference for interpreting the calibrations of the six CR questions. Besides the MC items, the CR questions range in difficulty from 0.38 for CR item 6 to 1.78 logits for CR item 2, with a mean difficulty value of 1.00 (SD = 0.67); average ratings range from 0.70 for CR item 2, which is most difficult, to 2.10 for CR items 6 and 1, which are easiest. As indicated by mean Infit and Outfit *MSE* statistics near the expected value of 1.00 with a standard deviation near 0.20, all six CR questions are functioning as intended by the rating scale MFR model for items.

Rating Scale Structure. Table 7 summarizes operational rater use of the rating scale for this sample, as described by the Rating Scale formulation of the MFR model. Category coefficient locations on the logit scale, estimated using this model, are displayed visually in Panel A of Figure 3. Inspection of the rating scale structure in terms of the quantitative guidelines listed in Table 2 indicates that the rating scale categories for the CR questions on the AP Statistics Exam are generally cooperating to produce meaningful measures of student statistics achievement. In terms of the first guideline (Directionality), the rating categories appear to be aligned with the latent variable, as indicated by a close match between observed and expected average measures. The alignment between observed and model-expected ratings is displayed graphically in Figure 4. Furthermore, the monotonic increase in category average measures indicates adherence to Guideline 2 and suggests agreement between the observed and intended ordering of categories.

In terms of category usage and rating distributions (Guidelines 3 and 4), category usage frequencies and percentages indicate that there is a generally good spread of ratings across scale categories. Furthermore, Outfit *MSE* statistics for all five rating scale categories are near their expected value of 1.00 (Guideline 5).

Evidence for adherence to Guideline 6 is indicated by the monotonic increase of category coefficient orders on the latent variable, with the lowest location observed for the threshold between the 0 and 1 category (−1.32 logits), and the highest location observed between categories 3 and 4 (1.91 logits). In contrast, an examination of the category coefficient location differences in terms of Guideline 7 indicates that the locations of rating scale categories may not be distinctive. Specifically, the difference between the location of the 0/1 category coefficient (−1.32 logits) and the 1/2 category coefficient location (−1.03 logits) has an absolute value of |0.29| logits, which is below the minimum difference of |1.40| logits given in Linacre (1999, 2002).

Visual Displays for Rating Scale Category Functioning. Visual displays for examining rating scale category functioning for the rating scale MFR model are provided in Figure 5. Because Model I imposes the same rating scale structure across all items, the visual displays in this figure apply across all six CR questions in the AP Statistics Exam. Panel A displays the category probability curves; five distinct peaks indicate that each rating scale category is most probable along a range of locations on the latent variable (x-axis), suggesting empirical observations in line with intended category ordering.

Panels B and C demonstrate a close match between the model-expected and empirically observed relationship between ratings and latent variable locations; however, some misfit beyond the 95% confidence interval occurs around rubric score categories 3 and 4 , which may imply different interpretations of ratings for students who earn ratings in these categories on CR questions.

Panels D and E, which display the cumulative and conditional probabilities for rating scale categories, respectively, indicate general adherence to rating scale guidelines, but differences in distance between curves in both of these plots reflect differences in category usefulness for distinguishing between intervals on the latent variable.

Finally, the item information display (Panel F) and category information display (Panel G) demonstrate that the information provided by the CR questions in the AP Statistics Exam varies over different locations on the latent variable, and is most informative around the 0.00 logit measure, at which Category 3 provides the most information.

Model II

Variable Map. The variable map shown in Figure 2 represents calibrations from Model II for students, CR questions, raters, and items. The interpretation of this variable map is similar to that for Figure 1. However, because Model II does not impose a fixed rating scale across the six CR questions, a different rating scale structure is shown for each item. As can be seen in the figure, the structure of these six rating scales is not equivalent. Specific information about each of these six scales is discussed in detail in the interpretation of Figures 4–9.

Summary Statistics. Table 8 provides summary statistics from the Model II Facets analyses for the students, MC items, CR questions, and raters examined in this study. As observed for Model I, the overall differences between students, MC items, CR questions, and raters are significant ($p < 0.05$), indicating a spread of the elements within each of the facets across the latent variable (statistics achievement). Similar values to Model I are observed in the Model II results for Infit and Outfit *MSE*, and suggest good fit to the partial-credit MFR model.

Calibration of CR Questions. Table 9 presents the results of CR question calibration for the AP Statistics Exam data under Model II. Once again, the MC items are anchored at 0.00 in order to provide a frame of reference for the interpretation of the locations of the six CR questions. Question locations are slightly lower for this model, and range from −0.30 logits for Question 6, which is easiest, to 0.88 logits for Question 2, which is most difficult. Average ratings range from 0.82 for Question 2 to 2.16 for Question 6. The Infit and Outfit *MSE* statistics vary slightly more from their expected value of 1.00 than was observed in the Model I findings; this observation is likely related to the nature of the partial-credit MFR model, which allows a different operational rating scale structure for each mode. It is interesting to note that the relative difficulty of the MC items changes between these two models. When calibrated under Model I, all of the CR questions are more difficult than the MC items. In contrast, CR Questions 6 and 1 are easier than the MC items under Model II.

Rating Scale Structure. Table 10 summarizes the structure of the rating scales for the six CR questions on the AP Statistics Exam. In terms of the first guideline (Directionality), the rating categories appear to be aligned with the latent variable, as indicated by a close match between observed and expected average measures for all questions except Question 4; the alignment between observed and model-expected ratings is displayed in Figure 6. Along the same lines, the rating scale categories demonstrate monotonic increase in category average measures, indicating adherence to Guideline 2 and suggesting agreement between the observed and intended ordering of categories for all questions except Question 4.

In terms of category usage and rating distributions (Guidelines 3 and 4), category usage frequencies and percentages indicate that there is a generally good spread of ratings across scale categories for all questions except 2 and 3. These two questions have fewer than 10 observations for the highest category (4: Complete Response). Infrequent observations within the highest category for these two CR questions are likely related to the fact that they are the most difficult of the CR questions in the AP Statistics Exam.

Outfit *MSE* statistics across the rating scale categories for all six CR questions are below 2.00 — following the requirements of Guideline 5. Furthermore, evidence for adherence to Guideline 6 is indicated by the monotonic increase of category coefficient orders on the latent variable for all CR questions; the lowest category coefficient location is observed for the threshold between the 0 and 1 category, and the highest location is observed between categories 3 and 4.

An examination of the category coefficient location differences in terms of Guideline 7 suggests that the locations of rating scale categories are not distinctive. Specifically, the difference between at least two of the locations of category coefficients have an absolute value below the minimum difference of 1.40 logits given in Linacre (1999, 2002) for all six CR questions. It is interesting to note that this category distinction guideline is violated between all category coefficient locations for Question 4, which was also identified as problematic by Guidelines 1 and 2.

Visual Displays for Rating Scale Category Functioning. Visual displays for examining rating scale category functioning for the partial-credit MFR model are provided in Panel B of Figure 3 and in Figures 6–12. Model II allows for a different rating scale structure within each CR question. As shown in Panel B of Figure 3, the results from analyses using the partial-credit MFR model do not support the hypothesis that the AP Statistics Exam rating scale is functioning consistently across the six CR questions, indicated by different category coefficient locations for each item. In this section, an overview of the rating scale category functioning is provided across all six CR questions, with an emphasis on unexpected elements in the graphical displays.

An examination of the rating scale category displays in Figures 7–12 reveals that the rating scale categories are functioning in a productive function across Panels A to G for only two of the six CR questions: CR Question 1 and CR Question 6. This finding is in line with a lack of significant problems detected by the quantitative indices of rating scale quality described earlier for these two CR questions. Along the same lines, Panels B and C reflect a generally good match between observed and empirical ICCs for all six CR questions, although there are occasional deviations beyond the 95% confidence intervals.

When compared across the six CR questions, the category response functions displayed in Panel A are problematic for CR Questions 2, 3, and 5. One category response function curve for each of these CR questions lacks modality for a range of values on the latent variable, suggesting that raters may not be using the non-modal categories as intended when scoring these CR questions.

Panels D and E display the cumulative and conditional probabilities for rating scale categories, and they do not demonstrate equal intervals on the x-axis for any of the CR questions. However, a lack of distinct categories in Panel E is particularly notable for CR Questions 2, 3, and 4. The overlapping conditional probability curves for these three CR questions suggest that two pairs of adjacent categories do not distinguish between students who are located in a particular range on the latent variable.

Panels F and G are graphical representations of the statistical information provided by rating scale categories along the latent variable. In general, the rating scale categories for this set of CR questions provide an overall peaked information curve (Panel F) and rating scale categories provide distinct information (Panel G) along a range of student locations on the latent variable. However, Questions 3 and 4 are problematic in terms of these two displays. In CR Question 3, item information is lost near the middle of the x-axis (around 0 logits), and the information provided by category 4 does not cover a distinct region on the x-axis. Along the same lines, Panel F for CR Question 5 indicates a narrow range of latent-variable locations for which high values of information are provided, and Panel G indicates that rating scale category 2 does not contribute to the information provided by this CR question.

Conclusions

In this section, findings are discussed as they apply to the two models used for analysis. Conclusions are provided as they relate to the three main research questions used to guide this study. Implications for research, theory, policy, and practice are discussed.

Model I: Rating Scale MFR Model

Model I is a rating scale MFR model, and it was used in this study to examine the spread of facet locations along the latent variable (statistics achievement). Findings from Model I analyses indicate significant differences ($p < 0.05$) within the groups of students and raters, and among the questions and MC items in the AP Statistics Exam. In terms of the first two research questions, these results suggest that rater severity and constructed-response item difficulty varies within the AP Statistics Exam.

Model I was also used to examine the quality of the rating scale structure used to score the six CR questions in terms of the quantitative rating scale category guidelines set forth by Linacre (1999, 2002), and a series of seven graphical displays from the Facets computer program (Linacre, 2010). Because Model I imposes a fixed rating scale structure on all of the CR questions, the quantitative and graphical indices of rating scale category quality for the AP Statistics Exam could be summarized across all six CR questions. In general, the quantitative descriptions of rating scale categories suggest adherence to the seven guidelines listed in Table 2. However, possible redundancy was observed between the first two rating scale categories, indicated by a small absolute value of the difference between rating scale category coefficient locations (Guideline 7). A similar story emerged in the examination of Model I results using the seven graphical displays from the Facets program (Linacre, 2010). All of the displays demonstrate appropriate rating scale category cooperation except for Panels D and E, which display cumulative and conditional probability curves for rating scale categories; these displays for Model I suggest a lack of distinction between intervals of student locations on the latent variable — a finding that reflects the violation of Guideline 7 by these rating scale data.

Model II: Partial-Credit MFR Model

Because Model II allows the rating scale structure to vary across each of the CR questions, it was possible to identify differences in the rating scale structure used to score each question on the AP Statistics Exam. In order to answer the third research question, quantitative indices from Linacre (1999, 2002) and Engelhard (2002), along with graphical displays from the Facets program (Linacre, 2010) were examined separately for each of the six AP Statistics CR questions. Overall, results from Model II analyses suggest that the structure of the rating scale is not comparable across the CR questions on the AP Statistics Exam. These differences are likely attributable to the fact that the scoring guidelines provide item-specific criteria that vary over each of these questions, despite the common rubric score labels.

Furthermore, findings highlight the differences in evidence for rating quality that are provided by each of the quantitative and graphical indices examined in this study. For example, the ratings assigned to CR Question 4 do not meet Guidelines 1 (Directionality), 2 (Monotonicity), or 7 (*Distinct* Category Coefficient Locations). However, Guidelines 3 (Category Usage), 4

(Distribution of Ratings Across Categories), 5 (Rating Scale Fit), and 6 (Category Coefficient Order) are not violated for CR Question 4. Only Guideline 7 provides a consistent description of rating quality across all six CR questions: Its violation by each rating scale was indicated by at least one instance of non-distinct category coefficient locations. Similarly, in Figure 9, Panel C (expected score and empirical ICC) suggests model-data fit within the 95% confidence bands for CR3. However, Panels F and G in the same figure indicate that the information provided by rating scale categories may not match model expectations for this question. Taken together, the set of quantitative indices and graphical displays described in this study as indices of rating quality suggest that multiple sources of evidence for rating quality are essential when interpreting rater-assigned scores.

Discussion and Summary

This study employs methodological tools based on Rasch Measurement Theory that have been used to examine the quality of ratings assigned in large-scale rater-mediated performance assessments. Two models (rating scale and partial-credit MFR models) were selected to demonstrate the application of Rasch-based rating quality indices to data from the AP Statistics Exam.

A key issue underlying this study is the interpretation of rater-assigned scores in performance assessment. When rating scale categories do not match the requirements for Rasch Theory as reflected in lack of adherence to the guidelines provided by Linacre (1999, 2002), then the interpretation of scores from CR questions may not reflect the expected and intended meaning of rating scale categories by the test developers. Methods for detecting rating scale idiosyncrasies based on Rasch Measurement Theory are a useful method for “flagging” CR questions, individual raters, or student performances for detailed qualitative interpretation, which may reveal contextual factors related to unexpected responses.

Discrepant rating scale application across the CR questions on the AP Statistics Exam has implications for the development of item-specific rubrics, rater training and monitoring for operational scoring, and the development of CR items. Because common performance descriptors are used across the CR questions on this exam, analyses using a partial-credit model can identify specific questions for which the community of meaning for the scoring rubric is not consistent among raters. As a result, revisions can be made to item-specific rubrics and rater training in order to establish a common understanding and application of the rating scale for these items. Finally, the construction of CR items can be informed by partial-credit findings. For example, issues related to distinct rating scale categories and category information could be used to develop items that provide more explicit opportunities to capture evidence of differences in AP Statistics Exam achievement within each rubric category.

The implications for this study are significant for both the development and interpretation of rating scales for large-scale rater-mediated performance assessments. Findings from this study suggest that a variety of methods can be used to evaluate rating quality based on an invariance framework. The observation that each of the Rasch-based quantitative and graphical indices of rating quality provides different information about the structure of a rating scale suggests that the interpretation of rater-assigned scores must be informed by the combined examination of these indices. Current methods for evaluating the quality of ratings in performance assessment are usually limited to quantitative measures of rater agreement as a method for explaining score variation across CR questions. In light of this study’s findings, detailed examination of rating scale functioning at the individual item level is essential for evaluating the quality of scores assigned to these items.

References

Andrich, D. A. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4), 581–594.

Andrich, D. A. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Andrich, D., de Jong, J.H.A.L., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In J. Rost & R. Langeheine (Eds.) *Applications of Latent Trait and Latent Class Models in the Social Sciences.* Munster, Germany: Waxmann Verlag Gmbh, 59–70.

College Board: Advanced Placement Program (2010). *AP Statistics 2010 scoring guidelines.* New York: The College Board.

College Board: Advanced Placement Program (2010). *Statistics course description.* New York: The College Board.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Eds.), *Large-Scale Assessment Programs for All Students: Development, Implementation, and Analysis* (pp. 261–287). Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken [Focus article]. *Measurement: Interdisciplinary Research and Perspectives 6*(3) 155–189.

Engelhard, G. (2009). Using item response theory and model-data fit to conceptualize differential item functioning for students with disabilities. *Educational and Psychological Measurement, 69*(4), 585–602.

Fisher, R. A. (1958). *Statistical methods for research workers.* New York: Hafner Publishing Co.

Huynh, H. (1994). On equivalence between a partial credit item and a set of independent Rasch binary items. *Psychometrika, 59*(1), 111–119.

Linacre, J. M. (1989). *Many-facet Rasch measurement.* Chicago: MESA press.

Linacre, J. M. (1997). Guidelines for rating scales. Retrieved from http://mesa.spc.uchicago.edu/rn2.htm

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, 3(2), 103–122.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85–106.

Linacre, J. M. (2010). Facets (Version 3) [Computer program]. Chicago: MESA Press.

Miao, J. & Odumade, O. (2011). *College Board Advanced Placement Program Examination, Statistics form 4GBP, reader reliability study.* Unpublished statistical report, Princeton, NJ: Educational Testing Service.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).

Table 1.

Theoretical and Empirical Rating Categories

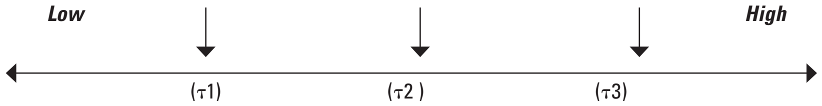
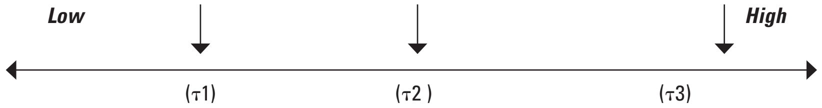
Panel A: Theoretical rating categories				
Ratings:	1	2	3	4
Labels:	Inadequate	Minimal	Good	Very Good
Theoretical mapping on latent variable				
Panel B: Empirical rating categories				
Ratings:	1	2	3	4
Labels:	Inadequate	Minimal	Good	Very Good
Empirical mapping on latent variable				

Table 2.

Guidelines for Examining Rating Scale Category Usage

Guidelines		Questions
1	Directionality	Does the usage of the rating categories align with latent variable?
2	Monotonicity	Do person locations on latent variable increase with rating categories?
3	Category usage	Are there sufficient observations per category?
4	Distribution of ratings	What is the distribution of ratings across categories?
5	Rating scale fit	How good is the fit of rating scale to Rasch model?
6	Category coefficient order	Do the locations of thresholds reflect the intended order of categories?
7	Category coefficient locations	Are the locations of thresholds distinctive?

Note: τ is defined as the category coefficient parameter.

Table 3.

AP Statistics Demographic Information

	Subgroups	Sample (*N* = 239)	
		N	%
Gender	Female	117	49.0
	Male	122	51.0
Race/Ethnicity	American Indian or Alaska Native	117	49.0
	Asian, Asian American, or Pacific Islander	3	1.3
	Black or African American	11	4.6
	Mexican or Mexican American	28	11.7
	Puerto Rican	1	0.4
	Other Hispanic, Latino, or Latin American	14	5.9
	White	159	66.5
	Other	10	4.2
	No Response	13	5.4
Best Language	English Only	191	79.9
	English and Another Language	37	15.5
	Another Language	3	1.3
	No Response	8	3.3

Table 4.

Model I (Rating Scale): Summary Statistics from Facets Analyses

	Students	MC Items	CR Questions	Raters
Measures				
M	0.25	0.00	1.00	0.00
SD	1.00	1.29	0.67	0.50
N	239	39	6	156
Infit *MSE*				
M	0.99	1.00	0.99	0.97
SD	0.29	0.08	0.15	0.42
Outfit *MSE*				
M	1.01	1.03	0.96	0.94
SD	0.39	0.20	0.15	0.40
Reliability of Separation	0.92	0.98	0.99	0.57
x^2 Statistic	2544.50*	1954.40*	1497.30*	319.90*
Degrees of Freedom	238	38	5	155

*$p < 0.05$

Table 5.

Model I (Rating Scale): Calibration of CR Questions Facet

	Measure	SE	Infit *MSE*	Outfit *MSE*	Mean Rating
MC items	0.00	0.03	1.00	0.20	0.50
6	0.38	0.06	0.82	0.83	2.10
1	0.66	0.06	1.18	1.19	2.10
5	1.12	0.06	1.05	1.01	1.39
4	1.48	0.06	1.05	1.01	1.30
3	1.57	0.06	0.76	0.74	1.18
2	1.78	0.07	1.06	0.93	0.70
M	1.00	0.05	0.99	0.96	1.30
SD	0.67	0.01	0.15	0.15	0.60

Table 6.

Model I Calibration of Rater Facet

Rater ID	Average Rating	Severity Measure (Logits)	SE	Infit *MSE*	Outfit *MSE*
71553	1.00	0.45	0.58	0.15	0.17
70652	2.75	-1.13	0.63	0.24	0.29
48168	0.95	0.75	0.28	0.33	0.34
46676	0.83	0.51	0.27	0.40	0.36
48236	0.87	0.44	0.27	0.39	0.37
71913	1.91	0.07	0.24	0.37	0.37
28890	1.05	0.24	0.28	0.48	0.44
69935	2.39	-0.43	0.25	0.43	0.45
76096	2.43	-0.69	0.25	0.43	0.45
45548	2.24	-0.41	0.29	0.46	0.47
88305	2.22	-0.09	0.40	0.44	0.47
21830	2.17	-0.27	0.24	0.49	0.48
74060	2.31	-0.65	0.33	0.49	0.48
75453	1.43	0.73	0.44	0.52	0.48
24731	1.39	-0.20	0.28	0.50	0.50
69729	2.14	-0.17	0.25	0.52	0.53
25870	1.10	0.01	0.27	0.57	0.55
88444	2.38	-0.08	0.24	0.52	0.55
95503	2.00	-0.18	0.29	0.58	0.55
16849	2.38	-0.08	0.24	0.52	0.56
26025	1.20	0.29	0.26	0.57	0.56
56503	0.95	0.31	0.26	0.60	0.57
23110	2.21	-0.44	0.17	0.59	0.60
31469	0.86	0.45	0.27	0.64	0.61
28448	1.32	0.50	0.26	0.73	0.62
28440	1.38	0.01	0.17	0.62	0.63
39347	1.29	-0.13	0.30	0.63	0.63
16853	1.21	0.48	0.24	0.68	0.64
20955	0.71	0.57	0.30	0.78	0.64
97943	1.00	0.34	0.32	0.74	0.65
19612	1.71	0.22	0.28	0.67	0.67
78382	1.54	0.09	0.30	0.66	0.67
27780	1.86	-0.04	0.18	0.71	0.69
76637	2.43	-0.69	0.25	0.70	0.69
31482	1.38	0.12	0.24	0.72	0.70

Note: Raters are ordered by Outfit *MSE*.

Table 6. (cont.)

Model I Calibration of Rater Facet

Rater ID	Average Rating	Severity Measure (Logits)	SE	Infit *MSE*	Outfit *MSE*
71908	1.50	-0.41	0.59	0.68	0.70
23116	1.00	0.40	0.26	0.69	0.71
24733	1.12	0.60	0.25	0.77	0.72
49571	1.04	0.14	0.25	0.75	0.72
79358	1.36	0.45	0.26	0.84	0.72
88872	1.00	0.22	0.26	0.73	0.72
93763	1.30	0.09	0.25	0.71	0.72
47176	1.52	0.04	0.26	0.79	0.73
78088	1.61	0.42	0.24	0.74	0.73
76090	2.08	-0.32	0.24	0.80	0.74
93596	1.05	0.42	0.31	0.77	0.74
56512	1.11	0.05	0.17	0.77	0.75
46689	0.74	0.90	0.28	0.85	0.76
81345	2.53	-0.19	0.29	0.81	0.76
82012	1.36	0.43	0.26	0.88	0.76
23111	1.08	0.34	0.25	0.75	0.77
96968	1.79	-0.01	0.27	0.81	0.77
23030	1.69	-0.22	0.17	0.80	0.78
98075	1.43	0.17	0.26	0.84	0.78
27786	0.59	0.84	0.34	0.92	0.79
36656	1.52	0.02	0.24	0.80	0.79
72100	2.25	-0.93	0.57	0.72	0.79
47179	2.38	-0.43	0.26	0.76	0.80
49518	1.83	0.00	0.23	0.78	0.80
72567	2.00	-0.61	0.78	0.80	0.80
28889	0.96	0.47	0.26	0.80	0.81
49551	0.67	0.86	0.33	0.73	0.81
70251	1.83	0.15	0.23	0.81	0.81
49216	0.57	0.66	0.31	1.05	0.82
56500	1.58	-0.16	0.24	0.84	0.82
61221	2.17	-0.11	0.24	0.80	0.82
20973	1.48	-0.08	0.18	0.82	0.83
26036	1.45	0.08	0.24	0.85	0.83
35411	1.38	-0.09	0.24	0.86	0.83
15563	1.67	-0.22	0.23	0.83	0.85

Note: Raters are ordered by Outfit *MSE*.

Table 6. (cont.)

Model I Calibration of Rater Facet

Rater ID	Average Rating	Severity Measure (Logits)	SE	Infit *MSE*	Outfit *MSE*
47178	1.65	-0.06	0.25	0.93	0.85
49515	2.62	-0.04	0.31	0.82	0.86
36663	1.00	0.57	0.27	0.95	0.87
83747	1.76	-0.21	0.24	0.91	0.88
20989	2.08	-0.15	0.19	0.91	0.89
20997	1.52	-0.25	0.26	0.94	0.89
28442	1.43	0.07	0.24	0.92	0.89
34630	1.19	-0.13	0.26	0.99	0.89
43823	1.36	0.12	0.17	0.93	0.89
34639	1.24	0.24	0.27	1.02	0.90
72483	1.54	0.19	0.34	0.94	0.90
79988	1.06	0.10	0.18	0.88	0.90
42550	2.00	-0.85	0.37	0.93	0.91
71020	1.78	0.24	0.25	0.92	0.91
53212	1.95	-0.31	0.25	0.93	0.92
72975	2.00	-0.04	0.46	0.98	0.92
26683	0.91	0.30	0.26	1.05	0.93
78701	1.67	-0.04	0.26	0.94	0.93
48233	1.17	0.31	0.19	0.98	0.94
95408	0.83	0.17	0.27	1.12	0.94
16850	1.09	0.11	0.25	1.02	0.95
68795	0.61	0.55	0.30	1.34	0.95
75116	1.56	-0.02	0.29	1.01	0.95
85222	0.95	0.69	0.28	0.78	0.96
64876	2.22	-0.57	0.25	1.02	0.97
27891	2.33	-0.58	0.25	0.98	0.99
47188	1.67	0.47	0.28	1.02	0.99
46675	1.86	-0.31	0.26	0.95	1.00
71925	2.29	-0.02	0.24	0.99	1.00
78699	2.25	-0.53	0.28	1.04	1.00
24736	0.57	1.09	0.35	0.83	1.01
38799	0.62	0.76	0.31	1.33	1.01
69612	1.70	0.28	0.25	0.97	1.01
25937	0.55	0.57	0.31	1.21	1.02
43908	2.00	-0.05	0.23	1.04	1.02

Note: Raters are ordered by Outfit *MSE*.

Table 6. (cont.)

Model I Calibration of Rater Facet

Rater ID	Average Rating	Severity Measure (Logits)	SE	Infit *MSE*	Outfit *MSE*
16851	1.78	-0.07	0.24	1.02	1.03
24091	1.61	0.00	0.25	1.14	1.03
26021	0.50	0.67	0.33	1.17	1.03
94624	1.67	0.56	0.23	1.01	1.03
34736	0.74	0.60	0.28	1.31	1.04
19215	2.61	-1.07	0.25	1.12	1.05
19632	1.83	0.34	0.23	1.03	1.06
31477	1.50	0.00	0.24	1.05	1.06
64861	1.08	0.20	0.25	1.27	1.06
23141	2.62	-1.05	0.26	1.10	1.07
26087	1.05	0.19	0.19	1.08	1.08
14361	2.38	-0.14	0.24	1.12	1.09
20982	1.62	-0.03	0.24	1.11	1.09
73056	0.96	0.30	0.26	1.15	1.10
38807	1.13	0.28	0.25	1.09	1.11
49547	2.24	-0.46	0.25	1.09	1.11
21082	2.26	-0.46	0.25	1.11	1.12
29788	2.00	-0.04	0.24	1.16	1.13
47177	0.33	1.13	0.36	0.96	1.13
87899	1.04	0.15	0.25	0.89	1.13
27889	0.65	0.48	0.29	0.98	1.14
30831	1.00	0.04	0.29	0.93	1.14
37719	1.25	-0.01	0.24	1.02	1.14
80616	2.04	0.07	0.24	1.17	1.14
21073	0.35	1.11	0.36	1.33	1.15
62615	1.11	0.18	0.18	1.01	1.15
26678	1.63	-0.41	0.27	1.18	1.16
30780	2.00	-0.05	0.23	1.16	1.17
49520	1.62	0.13	0.26	1.30	1.17
43909	2.00	-0.25	0.25	1.07	1.18
72881	2.50	0.04	0.27	1.27	1.21
19626	0.50	0.59	0.35	1.44	1.22
31483	1.17	0.11	0.24	1.30	1.22
20999	1.96	-0.23	0.24	1.22	1.24
31471	1.83	-0.15	0.24	1.31	1.24

Note: Raters are ordered by Outfit *MSE*.

Table 6. (cont.)

Model I Calibration of Rater Facet

Rater ID	Average Rating	Severity Measure (Logits)	SE	Infit *MSE*	Outfit *MSE*
15558	2.54	-0.70	0.25	1.24	1.25
15612	2.36	-0.54	0.26	1.40	1.25
26686	0.70	0.39	0.29	1.16	1.25
73405	1.91	-0.16	0.26	1.24	1.25
46428	2.61	-1.07	0.25	1.32	1.26
72033	0.62	0.68	0.50	1.10	1.26
37194	2.06	-0.10	0.28	1.31	1.27
71431	1.27	0.05	0.18	1.38	1.27
95415	1.83	-0.13	0.24	1.29	1.28
26031	2.37	-0.81	0.28	1.28	1.29
41021	1.13	0.56	0.21	1.44	1.30
83633	2.50	-0.47	0.25	1.30	1.30
86986	1.91	-0.14	0.25	1.37	1.30
26018	1.78	-0.08	0.25	1.52	1.35
26699	2.08	0.15	0.24	1.41	1.36
27782	2.09	-0.37	0.25	1.29	1.38
76091	1.40	-0.26	0.25	1.38	1.38
27797	1.75	0.59	0.23	1.36	1.40
30830	1.25	-0.05	0.25	1.61	1.42
47471	2.30	-0.67	0.24	1.46	1.42
78199	2.00	-0.08	0.50	1.35	1.42
42547	3.33	-1.27	0.51	1.63	1.44
27783	2.50	-0.64	0.25	1.55	1.45
48235	2.39	-0.79	0.25	1.59	1.50
71583	2.62	-0.78	0.43	1.74	1.51
26032	1.52	0.01	0.24	1.64	1.56
22036	2.79	-0.93	0.26	1.58	1.65
41022	2.41	-0.61	0.27	2.03	1.83
48237	1.95	-0.54	0.24	1.87	1.83
96974	2.12	-0.52	0.28	1.87	1.88
17708	2.91	-1.01	0.28	1.72	2.08

Note: Raters are ordered by Outfit *MSE*.

Table 7.

Model I (Rating Scale): Rating Scale Structure

Rating Scale Category	Label	Average Measure		Category Usage (%)	Outfit *MSE*	Category Coefficient Location	
		Observed	Expected			Logit Scale Location	|Difference|
0	Little to No Understanding	-2.01	-1.98	740 (27.92)	1.00		
1	Minimal Response	-1.16	-1.15	583 (22.00)	0.90	-1.32	
2	Developing Response	-0.38	-0.42	746 (28.15)	0.80	-1.03	0.29
3	Substantial Response	0.43	0.28	452 (17.06)	1.00	0.43	1.46
4	Complete Response	1.91	1.07	129 (4.87)	1.10	1.91	1.48

Table 8.

Model II (Partial Credit): Summary Statistics from Facets Analyses

	Students	MC Items	CR Questions	Rater
Measures				
M	0.24	0.00	0.28	0.00
SD	1.01	1.28	0.45	0.51
N	239	39	6	156
Infit *MSE*				
M	0.99	1.00	0.98	0.96
SD	0.29	0.08	0.09	0.40
Outfit *MSE*				
M	1.02	1.03	0.96	0.93
SD	0.40	0.20	0.11	0.40
Reliability of Separation	0.92	0.98	0.99	0.56
x^2 Statistic	2533.60*	2147.70*	437.20*	340.90*
Degrees of Freedom	238	38	5	155

* $p < 0.05$

Table 9.

Model II (Partial Credit): Calibration of MC Items and CR Questions Facet

	Measure	SE	Infit *MSE*	Outfit *MSE*	Mean Rating
6	-0.30	0.06	1.04	1.04	2.16
1	-0.26	0.04	1.08	1.11	2.24
MC Items	0.00	0.02	1.01	1.04	0.57
5	0.10	0.04	0.83	0.81	1.78
4	0.55	0.05	0.98	0.95	1.30
3	0.67	0.05	0.89	0.87	1.16
2	0.88	0.05	0.94	0.89	0.82
M	0.23	0.04	0.97	0.96	1.43
SD	0.47	0.01	0.09	0.11	0.65

Table 10.

Model II (Partial Credit): Rating Scale Structure

<table>
<tr><th colspan="7">CR Question 1</th></tr>
<tr><th rowspan="2">Rating Scale Category</th><th colspan="2">Average Measure</th><th rowspan="2">Category Usage (%)</th><th rowspan="2">Outfit <i>MSE</i></th><th colspan="2">Category Coefficient Location</th></tr>
<tr><th>Observed</th><th>Expected</th><th>Logit Scale Location</th><th>|Difference|</th></tr>
<tr><td>0</td><td>-1.17</td><td>-1.26</td><td>48 (10.48)</td><td>1.30</td><td></td><td></td></tr>
<tr><td>1</td><td>-0.46</td><td>-0.51</td><td>95 (20.74)</td><td>1.20</td><td>-1.81</td><td></td></tr>
<tr><td>2</td><td>0.11</td><td>0.10</td><td>125 (27.29)</td><td>0.80</td><td>-0.54</td><td>1.27</td></tr>
<tr><td>3</td><td>0.56</td><td>0.64</td><td>129 (28.17)</td><td>1.40</td><td>0.45</td><td>0.99</td></tr>
<tr><td>4</td><td>1.27</td><td>1.26</td><td>61 (13.32)</td><td>1.00</td><td>1.88</td><td>0.45</td></tr>
<tr><th colspan="7">CR Question 2</th></tr>
<tr><th rowspan="2">Rating Scale Category</th><th colspan="2">Average Measure</th><th rowspan="2">Category Usage (%)</th><th rowspan="2">Outfit <i>MSE</i></th><th colspan="2">Category Coefficient Location</th></tr>
<tr><th>Observed</th><th>Expected</th><th>Logit Scale Location</th><th>|Difference|</th></tr>
<tr><td>0</td><td>-2.29</td><td>-2.26</td><td>243 (56.12)</td><td>1.00</td><td></td><td></td></tr>
<tr><td>1</td><td>-1.51</td><td>-1.56</td><td>93 (21.48)</td><td>0.80</td><td>-1.46</td><td></td></tr>
<tr><td>2</td><td>-0.89</td><td>-0.97</td><td>65 (15.01)</td><td>0.80</td><td>-0.69</td><td>0.77</td></tr>
<tr><td>3</td><td>-0.58</td><td>-0.36</td><td>27 (6.24)</td><td>1.10</td><td>0.30</td><td>0.99</td></tr>
<tr><td>4</td><td>0.84</td><td>0.24</td><td>5 (1.15)</td><td>0.50</td><td>1.81</td><td>1.51</td></tr>
<tr><th colspan="7">CR Question 3</th></tr>
<tr><th rowspan="2">Rating Scale Category</th><th colspan="2">Average Measure</th><th rowspan="2">Category Usage (%)</th><th rowspan="2">Outfit <i>MSE</i></th><th colspan="2">Category Coefficient Location</th></tr>
<tr><th>Observed</th><th>Expected</th><th>Logit Scale Location</th><th>|Difference|</th></tr>
<tr><td>0</td><td>-2.37</td><td>-2.29</td><td>152 (33.93)</td><td>1.00</td><td></td><td></td></tr>
<tr><td>1</td><td>-1.51</td><td>-1.55</td><td>145 (32.37)</td><td>0.80</td><td>-2.18</td><td></td></tr>
<tr><td>2</td><td>-0.92</td><td>-0.93</td><td>134 (29.91)</td><td>0.80</td><td>-0.92</td><td>1.26</td></tr>
<tr><td>3</td><td>0.09</td><td>-0.27</td><td>17 (3.80)</td><td>0.70</td><td>1.07</td><td>1.99</td></tr>
<tr><td>4</td><td>0.24</td><td>0.56</td><td>4 (0.89)</td><td>1.30</td><td>2.02</td><td>0.95</td></tr>
</table>

Table 10. (cont.)

Model II (Partial Credit): Rating Scale Structure

CR Question 4

Rating Scale Category	Average Measure		Category Usage (%)	Outfit MSE	Category Coefficient Location	
	Observed	Expected			Logit Scale Location	|Difference|
0	-1.64	-1.68	144 (33.96)	1.10		
1	-1.27	-1.07	106 (25.00)	1.00	-1.52	
2	-0.48	-0.59	109 (25.71)	0.80	-0.67	0.85
3	0.08	-0.13	52 (12.24)	0.70	0.39	1.06
4	-0.07	0.53	13 (3.07)	1.30	1.77	1.38

CR Question 5

Rating Scale Category	Average Measure		Category Usage (%)	Outfit MSE	Category Coefficient Location	
	Observed	Expected			Logit Scale Location	|Difference|
0	-1.60	-1.47	127 (28.93)	0.70		
1	-0.89	-0.89	65 (14.81)	0.60	-1.30	
2	-0.28	-0.37	118 (26.88)	0.80	-0.78	0.52
3	0.29	0.13	101 (23.00)	0.80	0.15	0.93
4	0.41	0.77	28 (6.38)	1.2	1.85	1.70

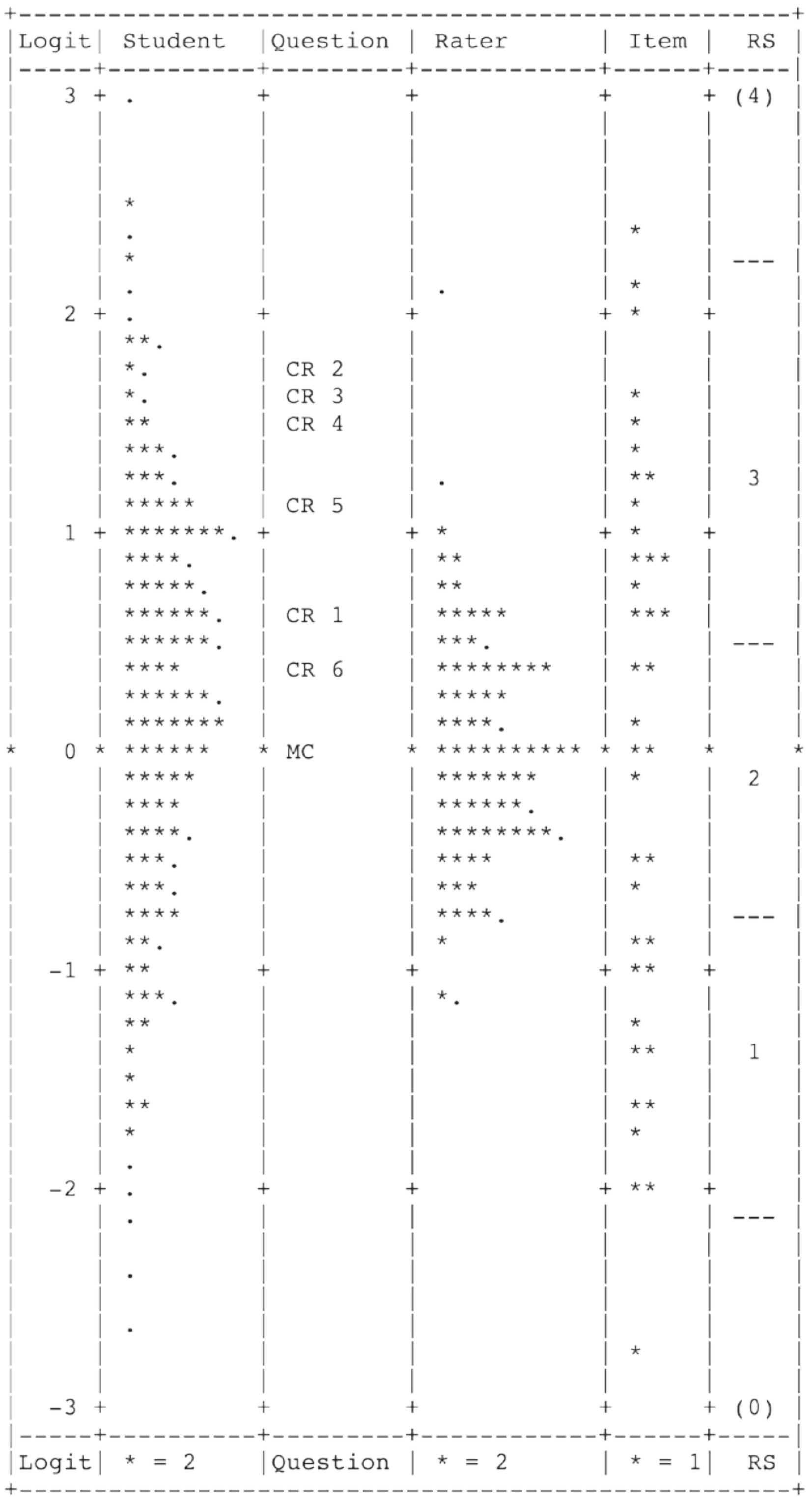
CR Question 6

Rating Scale Category	Average Measure		Category Usage (%)	Outfit MSE	Category Coefficient Location	
	Observed	Expected			Logit Scale Location	|Difference|
0	-1.42	-1.68	26 (5.86)	1.30		
1	-0.92	-0.87	79 (17.80)	0.90	-2.61	
2	-0.08	-0.10	195 (43.92)	1.00	-1.22	1.39
3	0.55	0.62	126 (28.38)	1.10	0.72	1.94
4	1.66	1.46	18 (4.05)	1.00	3.08	2.36

Note: Rating scale categories are labeled as follows: (0): Little to No Understanding; (1): Minimal Response; (2): Developing Response; (3): Substantial Response, (4): Complete Response.

Figure 1.

Model I (rating scale): Variable map



```
+----------------------------------------------------+
|Logit| Student  |Question | Rater      | Item |  RS |
|-----+----------+---------+------------+------+-----|
|   3 + .        +         +            +      + (4) |
|     |          |         |            |      |     |
|     |          |         |            |      |     |
|     |          |         |            |      |     |
|     | *        |         |            |      |     |
|     | .        |         |            | *    |     |
|     | *        |         |            |      | --- |
|     | .        |         | .          | *    |     |
|   2 + .        +         +            + *    +     |
|     | **.      |         |            |      |     |
|     | *.       | CR 2    |            |      |     |
|     | *.       | CR 3    |            | *    |     |
|     | **       | CR 4    |            | *    |     |
|     | ***.     |         |            | *    |     |
|     | ***.     |         | .          | **   |  3  |
|     | *****    | CR 5    |            | *    |     |
|   1 + *******. +         + *          + *    +     |
|     | ****.    |         | **         | ***  |     |
|     | *****.   |         | **         | *    |     |
|     | ******.  | CR 1    | *****      | ***  |     |
|     | ******.  |         | ***.       |      | --- |
|     | ****     | CR 6    | ********   | **   |     |
|     | ******.  |         | *****      |      |     |
|     | *******  |         | ****.      | *    |     |
*   0 * ******   * MC      * ********** * **   *     *
|     | *****    |         | *******    | *    |  2  |
|     | ****     |         | ******.    |      |     |
|     | ****.    |         | ********.  |      |     |
|     | ***.     |         | ****       | **   |     |
|     | ***.     |         | ***        | *    |     |
|     | ****     |         | ****.      |      | --- |
|     | **.      |         | *          | **   |     |
|  -1 + **       +         +            + **   +     |
|     | ***.     |         | *.         |      |     |
|     | **       |         |            | *    |     |
|     | *        |         |            | **   |  1  |
|     | *        |         |            |      |     |
|     | **       |         |            | **   |     |
|     | *        |         |            | *    |     |
|     | .        |         |            |      |     |
|  -2 + .        +         +            + **   +     |
|     | .        |         |            |      | --- |
|     |          |         |            |      |     |
|     | .        |         |            |      |     |
|     |          |         |            |      |     |
|     | .        |         |            |      |     |
|     |          |         |            | *    |     |
|     |          |         |            |      |     |
|  -3 +          +         +            +      + (0) |
|-----+----------+---------+------------+------+-----|
|Logit| * = 2    |Question | * = 2      | * = 1|  RS |
+----------------------------------------------------+
```

Figure 2.

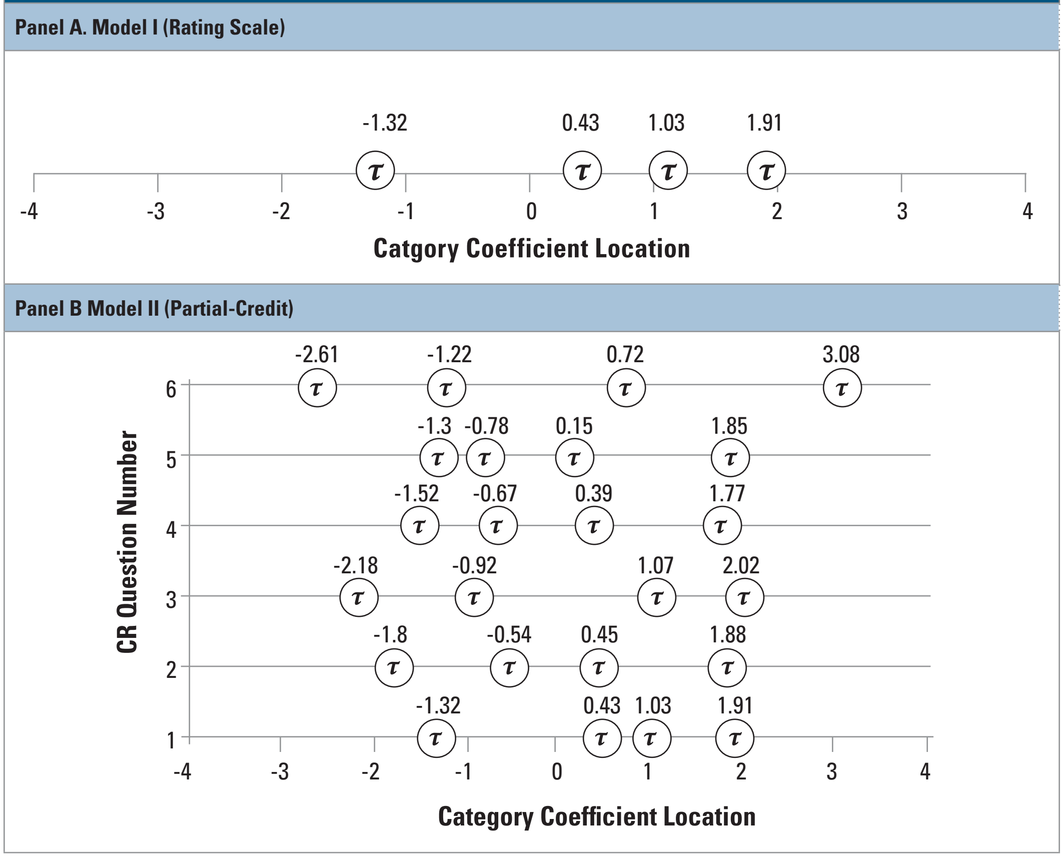
Model II (partial credit): Variable map

```
+----------------------------------------------------------------------------------+
|Logit| Student  |Question | Rater      | Item |PC.1 |PC.2 |PC.3 |PC.4 |PC.5 |PC.6 |
|-----+----------+---------+------------+------+-----+-----+-----+-----+-----+-----|
|     | .        |         |            |      | (4) | (4) | (4) | (4) | (4) | (4) |
|     |          |         |            |      |     |     |     |     |     |     |
|     |          |         |            |      |     |     |     |     |     | --- |
|   3 +          +         +            +      +     +     +     +     +     +     |
|     |          |         |            |      |     |     |     |     |     |     |
|     |          |         | .          |      |     |     |     |     |     |     |
|     |          |         |            |      |     |     |     |     |     |     |
|     | *        |         |            |      |     |     |     |     |     |     |
|     | .        |         |            | *    |     |     | --- |     |     |     |
|     | *.       |         |            |      |     |     |     |     |     |     |
|   2 + .        +         +            + **   + --- + --- +     + --- + --- +     |
|     | *        |         |            |      |     |     |     |     |     |  3  |
|     | **.      |         |            |      |     |     |     |     |     |     |
|     | **       |         |            | *    |     |     |  3  |     |     |     |
|     | **       |         |            | *    |     |     |     |     |     |     |
|     | ***      |         | .          | *    |     |     |     |     |     |     |
|     | *****    |         |            | **   |  3  |     |     |     |     |     |
|   1 + *******  +         + *.         + **   +     +  3  +     +  3  +     +     |
|     | *****.   |  CR 2   | *****      |      |     |     | --- |     |  3  |     |
|     | *******  |  CR 3   | *****      | ***  |     |     |     |     |     | --- |
|     | ******** |         | ******     | **   |     |     |     |     |     |     |
|     | *******  |  CR 4   | *******.   | ***  | --- |     |     | --- |     |     |
|     | ******   |         | *******    | **   |     | --- |     |     |     |     |
|     | ******   |  CR 5   | *******    |      |     |     |     |     | --- |     |
*   0 * *******. *   MC    * *********  * **   *     *     *  2  *     *     *     *
|     | *******  |         | *********. | *    |  2  |     |     |  2  |     |     |
|     | *****    |  CR 1   | *******.   | *    |     |  2  |     |     |  2  |  2  |
|     | ****.    |  CR 6   | ***        |      |     |     |     |     |     |     |
|     | ****     |         | ****       | **   |     |     |     |     |     |     |
|     | ****     |         | ***.       | *    | --- | --- |     | --- | --- |     |
|     | *****    |         |            |      |     |     | --- |     |     |     |
|  -1 + **       +         + **         + **   +     +     +     +     +     +     |
|     | **.      |         |            | **   |     |     |     |     |  1  | --- |
|     | ****.    |         |            |      |  1  |  1  |     |  1  |     |     |
|     | *        |         |            | ***  |     |     |     |     |     |     |
|     | *        |         |            |      |     |     |  1  |     |     |     |
|     | *.       |         |            | **   |     |     |     |     |     |     |
|     | *.       |         |            | *    |     | --- |     |     | --- |     |
|  -2 +          +         +            + *    +     +     +     + --- +     +  1  |
|     | *        |         |            | *    | --- |     |     |     |     |     |
|     | .        |         |            |      |     |     |     |     |     |     |
|     |          |         |            |      |     |     |     |     |     |     |
|     | .        |         |            |      |     |     | --- |     |     |     |
|     | .        |         |            |      |     |     |     |     |     |     |
|     |          |         |            | *    |     |     |     |     |     |     |
|  -3 +          +         +            +      +     +     +     +     +     + --- |
|     |          |         |            |      |     |     |     |     |     |     |
|     |          |         |            |      |     |     |     |     |     |     |
|     |          |         |            |      |     |     |     |     |     |     |
|     |          |         |            |      | (0) | (0) | (0) | (0) | (0) | (0) |
|-----+----------+---------+------------+------+-----+-----+-----+-----+-----+-----|
|Logit| * = 2    |Question | * = 2      | * = 1|PC.1 |PC.2 |PC.3 |PC.4 |PC.5 |PC.6 |
```

Note: The columns labeled “PC.k” represent the rating scale structure for CR question *k*, using the partial-credit MFR Model.

Figure 3.

Category coefficient locations



Panel A. Model I (Rating Scale)

Panel B Model II (Partial-Credit)

Figure 4.

Model I (rating scale): Observed and expected ratings

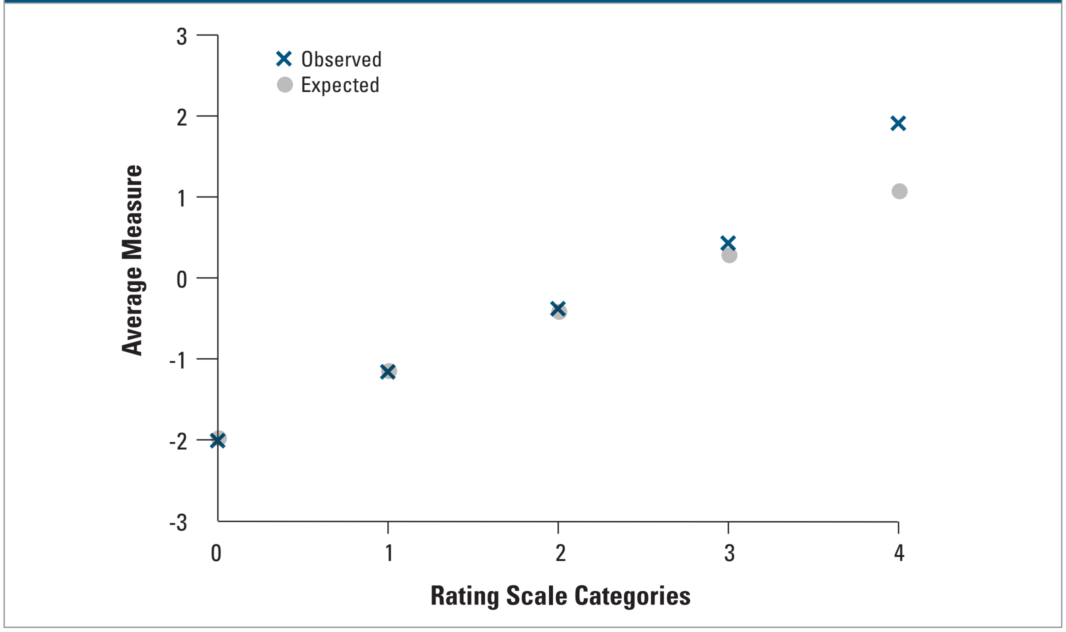
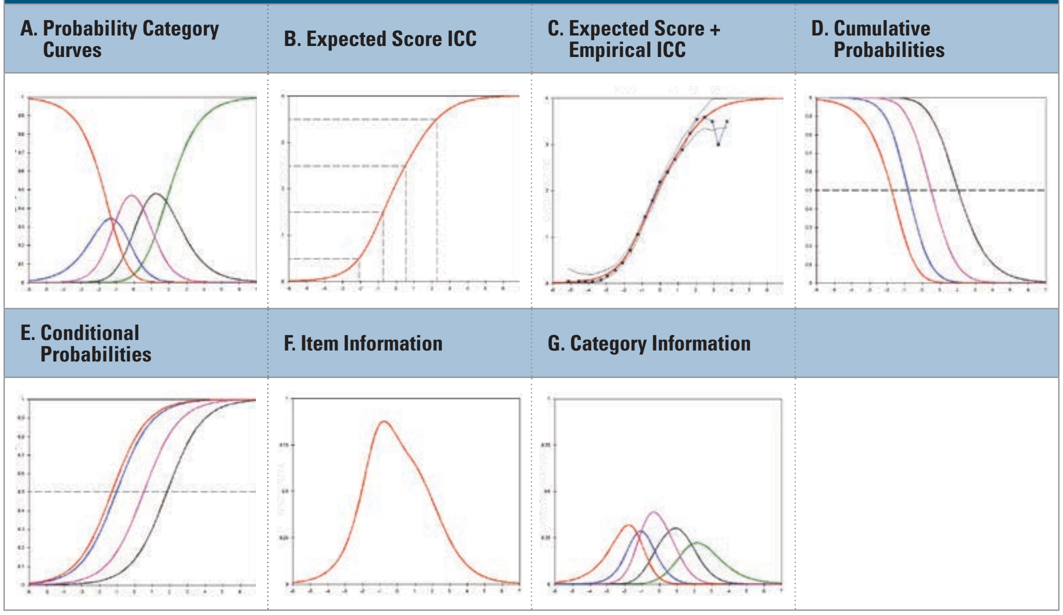


Figure 5.

Model I (rating scale): Question rating scale quality displays


A. Probability Category Curves

B. Expected Score ICC

C. Expected Score + Empirical ICC

D. Cumulative Probabilities

E. Conditional Probabilities

F. Item Information

G. Category Information

Figure 6.

Model II (partial credit): Observed and expected ratings

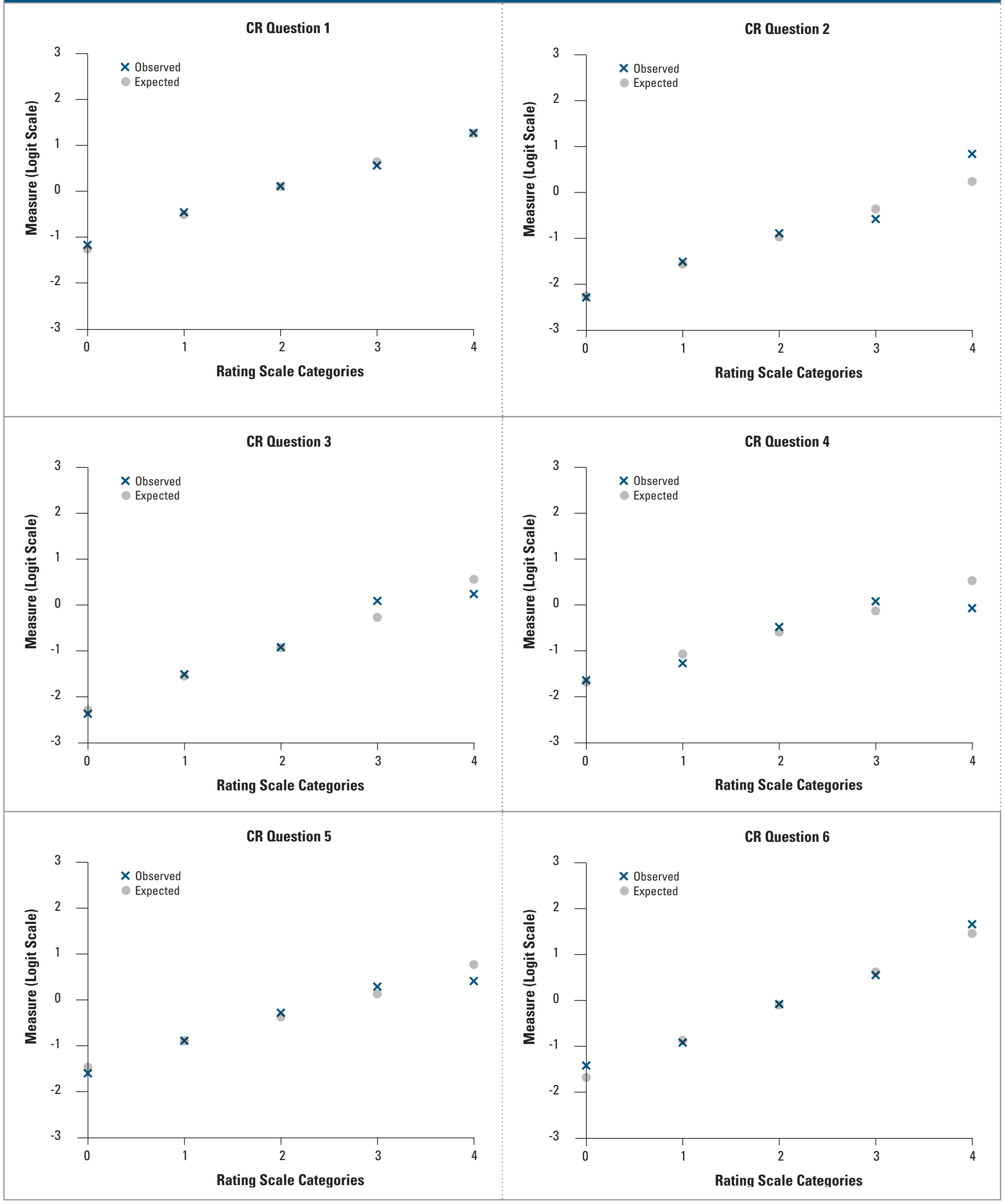
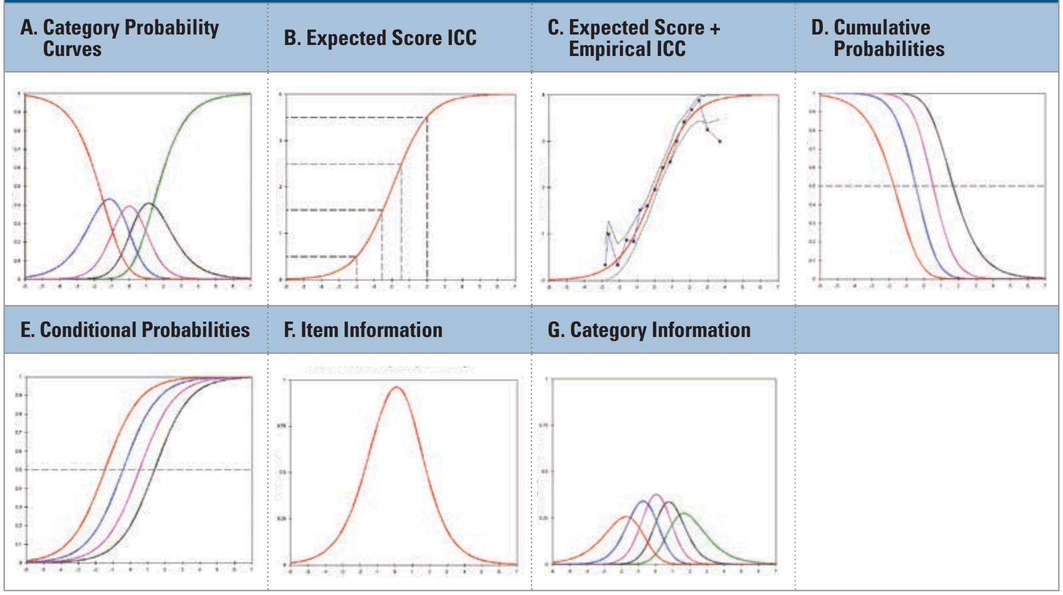


Figure 7.

Model II (partial credit), question 1



A. Category Probability Curves

B. Expected Score ICC

C. Expected Score + Empirical ICC

D. Cumulative Probabilities

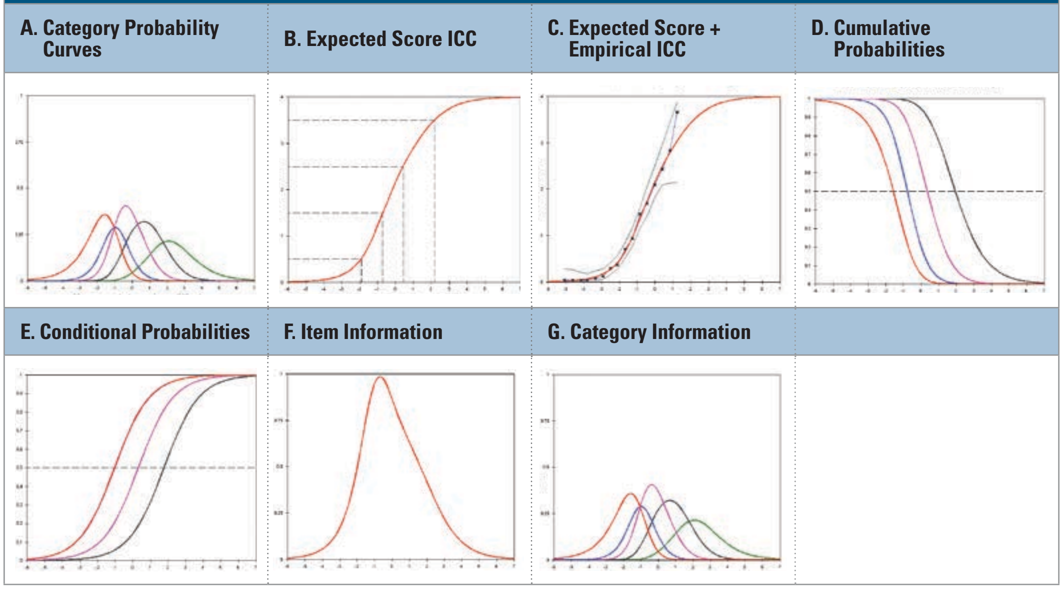
E. Conditional Probabilities

F. Item Information

G. Category Information

Figure 8.

Model II (partial credit MFR model), question 2



A. Category Probability Curves

B. Expected Score ICC

C. Expected Score + Empirical ICC

D. Cumulative Probabilities

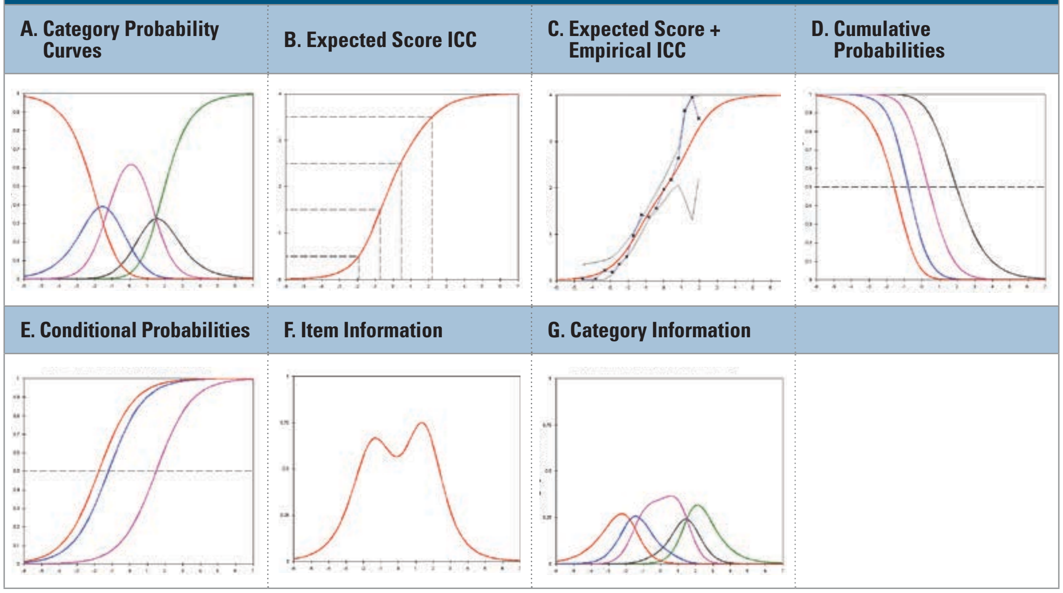
E. Conditional Probabilities

F. Item Information

G. Category Information

Figure 9.

Model II (partial credit MFR model), question 3



A. Category Probability Curves

B. Expected Score ICC

C. Expected Score + Empirical ICC

D. Cumulative Probabilities

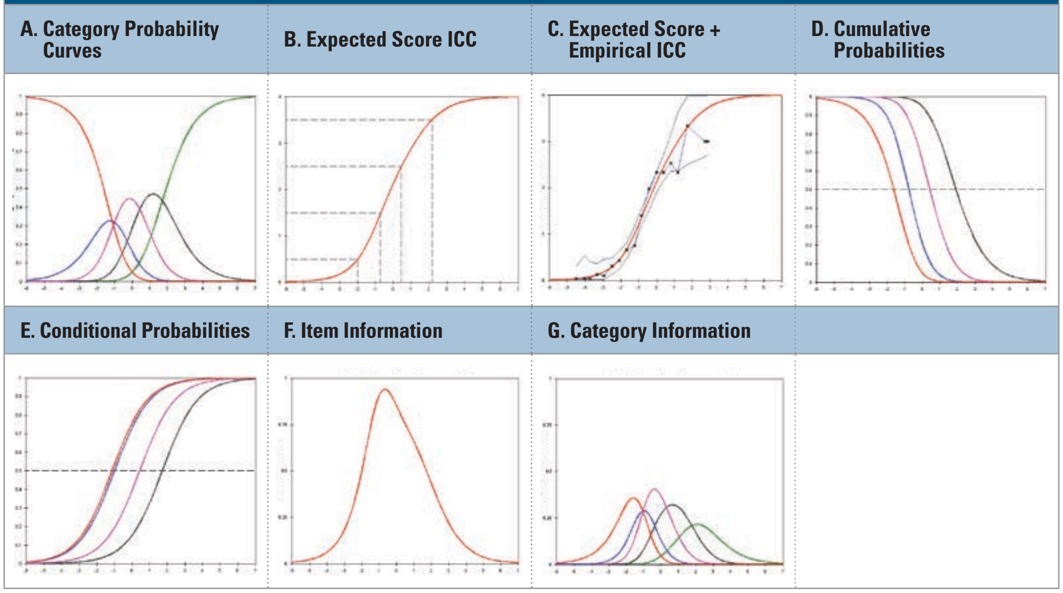
E. Conditional Probabilities

F. Item Information

G. Category Information

Figure 10.

Model II (partial credit MFR model), question 4



A. Category Probability Curves

B. Expected Score ICC

C. Expected Score + Empirical ICC

D. Cumulative Probabilities

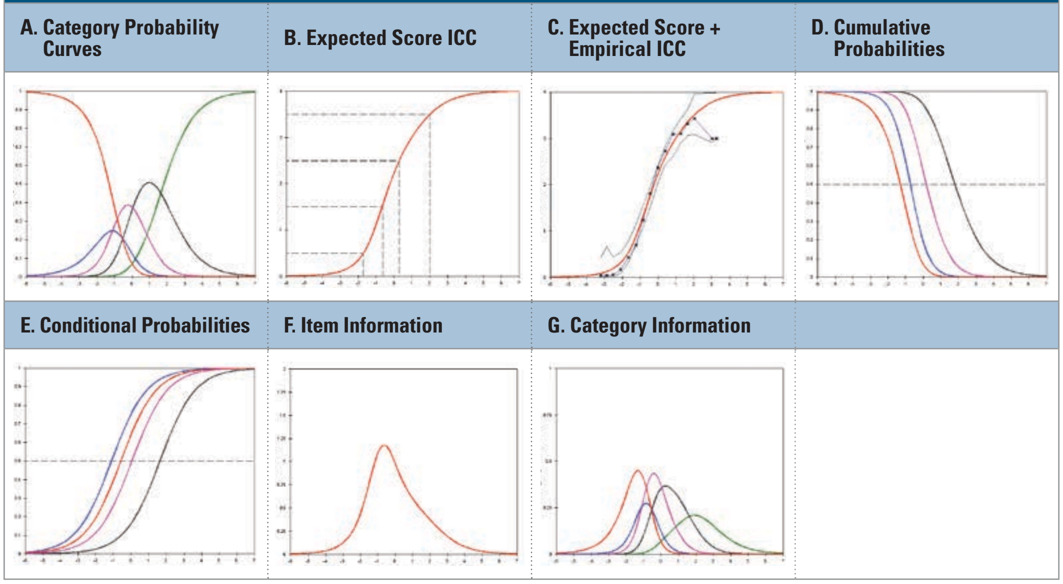
E. Conditional Probabilities

F. Item Information

G. Category Information

Figure 11.

Model II (partial credit MFR model), question 5



A. Category Probability Curves

B. Expected Score ICC

C. Expected Score + Empirical ICC

D. Cumulative Probabilities

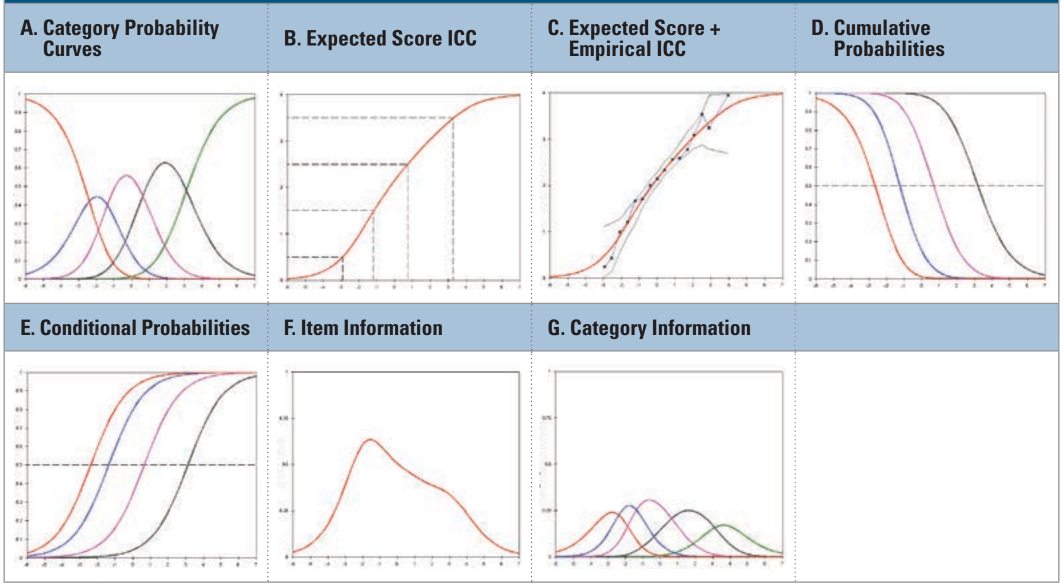
E. Conditional Probabilities

F. Item Information

G. Category Information

Figure 12.

Model II (partial credit MFR model), question 6



A. Category Probability Curves

B. Expected Score ICC

C. Expected Score + Empirical ICC

D. Cumulative Probabilities

E. Conditional Probabilities

F. Item Information

G. Category Information

Appendix

Table A1.

Category Descriptors

Score	Store Descriptor	Statistical Knowledge	Communication
		Identification of the important components of the problem; demonstration of the statistical concepts and techniques that result in a correct solution of the problem.	Explanation of what was done and why, along with a statement of conclusions drawn in context.
4	Complete Response	• Shows complete understanding of the problem's statistical components • Synthesizes a correct relationship among these components, perhaps with novelty and creativity • Uses appropriate and correctly executed statistical techniques • May have minor arithmetic errors but answers are still reasonable	• Provides a clear, organized, and complete explanation, using correct terminology, of what was done and why • States appropriate assumptions and caveats • Uses diagrams or plots when appropriate to aid in describing the solution • States an appropriate and complete conclusion in context
3	Substantial Response	• Shows substantial understanding of the problem's statistical components • Synthesizes a relationship among these components, perhaps with minor gaps • Uses appropriate statistical techniques • May have arithmetic errors but answers are still reasonable	• Provides a clear but not perfectly organized explanation, using correct terminology, of what was done and why, but explanation may be slightly incomplete • May miss necessary assumptions or caveats • Uses diagrams or plots when appropriate to aid in describing the solution • States a conclusion that follows from the analysis but may be somewhat incomplete

Table A1. (cont.)

Category Descriptors

Score	Store Descriptor	Statistical Knowledge	Communication
2	Developing Response	• Shows some understanding of the problem's statistical components • Shows little in the way of a relationship among these components • Uses some appropriate statistical techniques but misses or misuses others • May have arithmetic errors that result in unreasonable answers	• Provides some explanation of what was done, but explanation may be vague and difficult to interpret and terminology may be somewhat inappropriate • Uses diagrams in an incomplete or ineffective way, or diagrams may be missing • States a conclusion that is incomplete
1	Minimal Response	• Shows limited understanding of the problem's statistical components by failing to identify important components • Shows little ability to organize a solution and may use irrelevant information • Misuses or fails to use appropriate statistical techniques • Has arithmetic errors that result in unreasonable answers	• Provides minimal or unclear explanation of what was done or why it was done, and explanation may not match the presented solution • Fails to use diagrams or plots, or uses them incorrectly • States an incorrect conclusion or fails to state a conclusion
0	(Zero-Level Response)	Shows little to no understanding of statistical components	Provides no explanation of a legitimate strategy

Note: Adapted from *AP Statistics Course Description* (College Board, 2010).

Table A2.

Scoring Guidelines

Score	Score Descriptor	Scoring Guidelines
4	Complete Response	Both parts essentially correct
3	Substantial Response	One part essentially correct and one part partially correct
2	Developing Response	One part essentially correct and one part incorrect OR Both parts partially correct
1	Minimal Response	One part partially correct and one part incorrect
0	(Zero-Level Response)	Both parts incorrect

Note: Adapted from *AP Statistics Course Description* (College Board, 2010).

The Research department actively supports the College Board's mission by:

- Providing data-based solutions to important education problems and questions
- Applying scientific procedures and research to inform our work
- Designing and evaluating improvements to current assessments and developing new assessments as well as educational tools to ensure the highest technical standards
- Analyzing and resolving critical issues for all programs, including AP®, SAT®, PSAT/NMSQT®
- Publishing findings and presenting our work at key scientific and education conferences
- Generating new knowledge and forward-thinking ideas with a highly trained and credentialed staff

Our work focuses on the following areas

Admission

Alignment

Evaluation

Fairness

Measurement

Research

Trends

Validity

Follow us online: research.collegeboard.org

CollegeBoard