New York State Regents Examination in Physics

2011 Field Test Analysis, Equating Procedure, and Scaling of Operational Test Forms

Technical Report

Prepared for the New York State Education Department
by Pearson

December 2011

Copyright

Table of Contents

List of Tables

Section I: Introduction

PURPOSE

The purpose of this report is to document the psychometric work on the New York State Regents Examination in Physics in 2011. Specifically, contained within this report are procedures for and results of field test analysis, equating, and scaling of operational test forms that were conducted by Pearson. Information on test development can be found in the test design and development report for the New York State Regents Examination in Physics.

Section II: Field Test Analysis

In May 2011, field testing was conducted for the New York State Regents Examination in Physics to better understand the psychometric quality of the items. The results of this testing are used to help determine which items will be selected for use on operational tests.

Target student samples for participation in this testing were selected such that each would represent the student population expected to take the operational test. The Need/Resource Capacity Categories were used as variables in the sampling plan. See Table 1 for the seven Need/Resource Capacity Categories and their definitions.

Table 1. Need/Resource Capacity Category Definitions

Need/Resource Capacity (N/RC) Category	Definition
High N/RC Districts: New York City	New York City
Large Cities	Buffalo, Rochester, Syracuse, Yonkers
Urban-Suburban	Districts at or above the 70th percentile on the index with at least 100 students per square mile or enrollment greater than 2500
Rural	All districts at or above the 70th percentile with fewer than 50 students per square mile or enrollment of less than 2500
Average N/RC Districts	All districts between the 20th and 70th percentiles on the index
Low N/RC Districts	All districts below the 20th percentile on the index
Charter Schools	Each charter school is a district

The data collected from field testing were scored by two entities. The multiple-choice items were scored by the New York State Education Department, and the constructed-

response items were scored by Measurement Incorporated. Therefore, it was necessary to combine data files for data analysis. Both classical and item response theory analyses were conducted using the data to evaluate the quality of the test items.

FILE MERGING AND DATA CLEANUP

Field test forms contained multiple-choice and constructed-response item types. Response data were contained in two separate files. The multiple-choice data file contained 10,123 student records, and the constructed-response data file contained 9,864 student records. The two files were combined by merging the multiple-choice records and the constructed-response records by unique test booklet numbers. After the exclusion rules were applied, the resulting field test data file contained 8,456 records.

Multiple-choice response data were then compared to the answer key. All item responses not matching the answer key were assigned scores of 0. The responses matching the answer key were assigned scores of 1. With respect to the constructed-response items, scores from 0 to the maximum point value available for each tested item were kept while out-of-range values were assigned scores of 0. For IRT calibrations, blanks (i.e., missing data) were assigned scores of 0 to be consistent with how operational test items are scored.

The final data file contained both the scored and unscored student responses. Unscored data were used to calculate the percentage of students who selected the various answer choices for the multiple-choice items or the percentage of students who received the range of possible raw score points for the constructed-response items. Thus, the frequency of students leaving items blank can be calculated. The scored data were used for all other analyses.

CLASSICAL ANALYSIS

Classical Test Theory is based on the assumption that an observed test score x is composed of both true score t and error score e. This assumption is expressed as follows:

$$x = t + e$$

In other words, error is associated with measuring a student's true score. For example, the choice of test items or the administration conditions might influence student responses, making a student's observed score higher or lower than the student's true score. The error is considered random. After repeated administrations,

the mean of the error scores is virtually zero. Thus, a student's observed score is expected to equal his or her true score. This expectation is expressed as follows:

$$E(x) = t$$

Using a Classical Test Theory framework, field test data can be analyzed to provide information about the quality of test items. Item difficulties, point-biserial correlations, reliability estimates, and various statistics related to rater agreement have been calculated and are summarized in the following section.

Item Difficulty

Item difficulty is an indication of students' performance on a specific item. Because this examination contains polytomous items, item means are not appropriate for comparing difficulty across items. Instead, weighted item means were calculated by dividing an item's mean by the maximum points possible for that item.

For multiple-choice items, the item difficulty is the proportion of students who answer an item correctly. If 90% of the student responses to a multiple-choice item are correct, then this item is considered easier than a multiple-choice item with correct responses by 30% of the students.

Point-Biserial Correlation

The point-biserial correlation is another classical statistic that can be used to evaluate items. For multiple-choice items, it is the correlation between students' performance on a given item (correct or incorrect) and overall performance scores. This statistic is used to evaluate how well an item identifies students who understand the concept being measured, and can be generalized for constructed-response items. The possible range for the point-biserial correlation is −1 to 1, with higher values being more desirable.

Table 2 presents a summary of the classical item analysis for each of the field test forms. The first three columns identify the form number, the number of students who took each form, and the number of items on each field test form, respectively. The remaining columns are divided into two sections (i.e., item difficulty and point-biserial correlations). Recall that for constructed-response items, item means were divided by the maximum number of points possible in order to place them in the same metric as the multiple-choice items. No item difficulties were below 0.90. With respect to the point-biserial correlations, few of these correlations fell below 0.25.

Table 2. Classical Item Analysis

			Item Difficulty			Point-Biserial		
Form	N-Count	No. of Items	<0.50	0.50 to 0.90	>0.90	<0.25	0.25 to 0.50	>0.50
831	775	20	10	10	0	1	15	4
832	770	19	7	12	0	1	14	4
833	791	19	8	11	0	3	10	6
834	774	19	5	14	0	1	13	5
835	778	19	6	13	0	1	11	7
836	767	19	8	11	0	0	15	4
837	752	20	6	14	0	1	13	6
838	766	20	12	7	0	1	14	4
839	757	20	6	14	0	1	11	8
840	756	20	11	9	0	0	12	8
841	770	16	5	11	0	1	9	6

* For some forms, the item counts in the “Item Difficulty” and “Point-Biserial” columns may not sum to the value in the “No. of Items” column due to DNS (do not score) items.

In addition to the summary information provided in Table 2, all of the classical item statistics are provided in Appendix A. “Max” is the maximum number of possible points. “N-Count” refers to the number of student records in the analysis. “Alpha” contains the internal consistency statistics discussed below. For multiple-choice items, “B” represents the proportion of students who left the item blank, and “M1” through “M4” are the proportions of students who selected each of the four answer choices. For constructed-response items, “B” represents the proportion of students who left the item blank, and “M0” through “M2” are the proportions of students who received scores 0 through 2. “Mean” is the average of the scores received by the students. The final column contains the point-biserial correlation for each item. There are some instances of items missing statistics; this occurs when an item was not scored.

Test Reliability

Classical analysis can also be used to measure the reliability of the test. Reliability is the consistency of the results obtained from a measurement with respect to time or among items or subjects that constitute a test. As such, test reliability can be estimated in a variety of ways. Internal consistency indices are a measure of how consistently examinees respond to items within a test. Two factors influence estimates of internal consistency: test length and homogeneity of items. In general, the more items on the examination the higher the reliability, and the more similar the items are the higher the reliability.

Cronbach's α (alpha) (Cronbach, 1951) has an important use as a measure of the internal consistency of a test. This formula is the extension of an earlier version, the Kuder-Richardson Formula 20 (KR-20), which is the equivalent for dichotomous items.

Table 3 contains the internal consistency statistics for all of the field test forms. These statistics ranged from 0.71 to 0.80 and are based solely on the items in the individual field test forms. It is expected that these statistics associated with the operational tests would be greater because there are more items on the operational test forms.

Table 3. Test and Scoring Reliability

Form Number	Test Reliability	Scoring Reliability
831	0.78	0.91
832	0.76	0.84
833	0.71	0.91
834	0.77	0.95
835	0.80	0.93
836	0.75	0.90
837	0.77	0.88
838	0.73	0.91
839	0.79	0.89
840	0.80	0.95
841	0.71	0.94

Scoring Reliability

One concern with constructed-response items is the reliability of the scoring process (i.e., consistency of the score assignment). Constructed-response items must be read by scorers who assign scores based on a comparison between the rubric and student responses. Consistency in the way scores are assigned is a critical part of the reliability of the assessment. To measure this consistency, 10% of the test booklets are scored a second time (i.e., second read scores) and compared to the original set of scores (i.e., first read scores).

As an overall measure of scoring reliability, the Pearson Correlation Coefficient between the first and second scores for each of the constructed-response items was computed. This statistic is often used as an overall indicator of scoring reliability, and generally ranges from 0 to near 1. Table 3 contains the results from these analyses in the column headed “Scoring Reliability.” The correlations ranged from 0.84 to 0.95, indicating high scoring reliability.

Inter-rater Agreement

For each constructed-response item, the difference between the first and second reads was computed. When examining inter-rater agreement statistics, it should be kept in mind that the maximum number of points per item varies, as shown in the "Score Points" column of the following tables.

Table 4 contains the proportion of occurrence of these differences for each item. The majority of differences between the first read and second read were 0.

Table 4. Point Differences Between First and Second Reads

			Difference (First Read minus Second Read)				
Form	**Item**	**Score Points**	**−2**	**−1**	**0**	**1**	**2**
831	13	1	0.00	0.04	0.91	0.05	0.00
831	14	1	0.00	0.03	0.95	0.02	0.00
831	15	1	0.00	0.00	0.95	0.05	0.00
831	16	1	0.00	0.03	0.95	0.02	0.00
831	17	1	0.00	0.00	0.99	0.01	0.00
831	18	1	0.00	0.01	0.98	0.02	0.00
831	19	1	0.00	0.03	0.96	0.01	0.00
831	20	1	0.00	0.04	0.95	0.02	0.00
832	12	1	0.00	0.04	0.91	0.06	0.00
832	13	1	0.00	0.05	0.92	0.03	0.00
832	14	1	0.00	0.03	0.94	0.03	0.00
832	15	1	0.00	0.04	0.95	0.01	0.00
832	16	1	0.00	0.01	0.98	0.02	0.00
832	17	1	0.00	0.05	0.91	0.04	0.00
832	18	1	0.00	0.04	0.92	0.04	0.00
832	19	1	0.00	0.07	0.85	0.08	0.00
833	12	1	0.00	0.02	0.98	0.01	0.00
833	13	1	0.00	0.02	0.96	0.02	0.00
833	14	1	0.00	0.05	0.92	0.03	0.00
833	15	1	0.00	0.00	0.99	0.01	0.00
833	16	1	0.00	0.03	0.97	0.00	0.00
833	17	1	0.00	0.04	0.92	0.05	0.00
833	18	1	0.00	0.04	0.92	0.05	0.00
833	19	1	0.00	0.01	0.99	0.00	0.00

Table 4. Point Differences Between First and Second Reads *(continued)*

			Difference (First Read minus Second Read)				
Form	**Item**	**Score Points**	**−2**	**−1**	**0**	**1**	**2**
834	12	1	0.00	0.01	0.99	0.00	0.00
834	13	1	0.00	0.02	0.96	0.02	0.00
834	14	1	0.00	0.00	1.00	0.00	0.00
834	15	1	0.00	0.00	0.99	0.01	0.00
834	16	1	0.00	0.06	0.91	0.03	0.00
834	17	1	0.00	0.02	0.93	0.05	0.00
834	18	1	0.00	0.00	1.00	0.00	0.00
834	19	1	0.00	0.01	0.99	0.00	0.00
835	12	1	0.00	0.02	0.95	0.02	0.00
835	13	1	0.00	0.03	0.96	0.01	0.00
835	14	1	0.00	0.03	0.93	0.04	0.00
835	15	1	0.00	0.01	0.99	0.00	0.00
835	16	1	0.00	0.01	0.98	0.02	0.00
835	17	1	0.00	0.02	0.96	0.02	0.00
835	18	1	0.00	0.02	0.96	0.02	0.00
835	19	1	0.00	0.01	0.98	0.02	0.00
836	12	1	0.00	0.05	0.93	0.02	0.00
836	13	1	0.00	0.02	0.98	0.01	0.00
836	14	1	0.00	0.00	0.99	0.01	0.00
836	15	1	0.00	0.03	0.92	0.05	0.00
836	16	1	0.00	0.06	0.88	0.06	0.00
836	17	1	0.00	0.01	0.98	0.02	0.00
836	18	1	0.00	0.01	0.98	0.01	0.00
836	19	1	0.00	0.02	0.95	0.03	0.00
837	12	1	0.00	0.03	0.97	0.00	0.00
837	13	1	0.00	0.03	0.95	0.02	0.00
837	14	1	0.00	0.05	0.92	0.03	0.00
837	15	1	0.00	0.04	0.93	0.03	0.00
837	16	1	0.00	0.03	0.93	0.04	0.00
837	17	1	0.00	0.03	0.93	0.04	0.00
837	18	1	0.00	0.06	0.91	0.03	0.00
837	19	1	0.00	0.01	0.98	0.02	0.00

Table 4. Point Differences Between First and Second Reads *(continued)*

			Difference (First Read minus Second Read)				
Form	**Item**	**Score Points**	**−2**	**−1**	**0**	**1**	**2**
837	20	1	0.00	0.01	0.93	0.06	0.00
838	12	1	0.00	0.02	0.94	0.03	0.00
838	13	1	0.00	0.04	0.92	0.04	0.00
838	14	1	0.00	0.00	1.00	0.00	0.00
838	16	1	0.00	0.01	0.95	0.04	0.00
838	17	1	0.00	0.02	0.97	0.02	0.00
838	18	1	0.00	0.03	0.94	0.03	0.00
838	19	1	0.00	0.02	0.95	0.04	0.00
838	20	1	0.00	0.01	0.97	0.02	0.00
839	12	1	0.00	0.01	0.98	0.02	0.00
839	13	1	0.00	0.02	0.93	0.06	0.00
839	14	1	0.00	0.03	0.97	0.00	0.00
839	15	1	0.00	0.05	0.93	0.02	0.00
839	16	1	0.00	0.04	0.92	0.05	0.00
839	17	1	0.00	0.03	0.92	0.05	0.00
839	18	1	0.00	0.00	0.98	0.02	0.00
839	19	1	0.00	0.04	0.95	0.02	0.00
839	20	1	0.00	0.06	0.92	0.02	0.00
840	13	1	0.00	0.01	0.98	0.01	0.00
840	14	1	0.00	0.01	0.98	0.01	0.00
840	15	1	0.00	0.04	0.94	0.02	0.00
840	16	1	0.00	0.02	0.98	0.01	0.00
840	17	1	0.00	0.03	0.94	0.03	0.00
840	18	1	0.00	0.00	1.00	0.00	0.00
840	19	1	0.00	0.02	0.98	0.00	0.00
840	20	1	0.00	0.00	0.99	0.01	0.00
841	13	2	0.00	0.04	0.91	0.05	0.01
841	14	1	0.00	0.01	0.98	0.02	0.00
841	15	2	0.00	0.06	0.88	0.05	0.00
841	16	2	0.00	0.02	0.95	0.04	0.00

Table 5 contains additional summary information regarding the first and second read scores. In the fifth column, the percent of exact matches between the first and second

read scores is provided. “Adj.” is the percentage of differences with a magnitude of 1. “Total” is the sum of the two prior columns and contains values between 99.2% and 100%. These values indicate a high degree of agreement.

Table 5. First and Second Read Descriptive Statistics and Agreement

				Agreement (%)			Raw Score Mean		Raw Score Standard Deviation			
Form	Item	Score Points	Total N-Count	Exact	Adj.	Total	First Read	Second Read	First Read	Second Read	Intraclass Correla-tion	Wt. Kappa
831	13	1	135	91.1	8.9	100.0	0.6	0.6	0.49	0.49	0.82	0.82
831	14	1	125	95.2	4.8	100.0	0.5	0.6	0.50	0.50	0.90	0.90
831	15	1	125	95.2	4.8	100.0	0.7	0.6	0.46	0.48	0.90	0.89
831	16	1	130	95.4	4.6	100.0	0.7	0.7	0.46	0.45	0.89	0.89
831	17	1	130	99.2	0.8	100.0	0.5	0.5	0.50	0.50	0.98	0.98
831	18	1	130	97.7	2.3	100.0	0.5	0.4	0.50	0.50	0.95	0.95
831	19	1	132	96.2	3.8	100.0	0.1	0.1	0.30	0.33	0.81	0.81
831	20	1	132	94.7	5.3	100.0	0.5	0.5	0.50	0.50	0.89	0.89
832	12	1	137	90.5	9.5	100.0	0.7	0.6	0.48	0.48	0.79	0.79
832	13	1	137	92.0	8.0	100.0	0.5	0.5	0.50	0.50	0.84	0.84
832	14	1	135	94.1	5.9	100.0	0.6	0.6	0.49	0.49	0.88	0.88
832	15	1	135	94.8	5.2	100.0	0.6	0.6	0.49	0.48	0.89	0.89
832	16	1	133	97.7	2.3	100.0	0.1	0.1	0.34	0.33	0.90	0.90
832	17	1	124	91.1	8.9	100.0	0.5	0.5	0.50	0.50	0.82	0.82
832	18	1	124	91.9	8.1	100.0	0.5	0.5	0.50	0.50	0.84	0.84
832	19	1	134	85.1	14.9	100.0	0.5	0.5	0.50	0.50	0.70	0.70
833	12	1	132	97.7	2.3	100.0	0.8	0.8	0.39	0.38	0.92	0.92
833	13	1	125	96.0	4.0	100.0	0.4	0.4	0.50	0.49	0.92	0.92
833	14	1	125	92.0	8.0	100.0	0.7	0.8	0.44	0.43	0.79	0.79
833	15	1	141	98.6	1.4	100.0	0.7	0.7	0.47	0.48	0.97	0.97
833	16	1	141	97.2	2.8	100.0	0.4	0.4	0.49	0.50	0.94	0.94
833	17	1	133	91.7	8.3	100.0	0.3	0.3	0.47	0.47	0.81	0.81
833	18	1	133	91.7	8.3	100.0	0.5	0.5	0.50	0.50	0.83	0.83
833	19	1	130	99.2	0.8	100.0	0.5	0.5	0.50	0.50	0.98	0.98
834	12	1	129	99.2	0.8	100.0	0.7	0.7	0.46	0.46	0.98	0.98
834	13	1	129	96.1	3.9	100.0	0.5	0.5	0.50	0.50	0.92	0.92
834	14	1	129	100.0	0.0	100.0	0.5	0.5	0.50	0.50	1.00	1.00
834	15	1	136	99.3	0.7	100.0	0.8	0.8	0.42	0.43	0.98	0.98
834	16	1	117	91.5	8.5	100.0	0.3	0.4	0.47	0.48	0.81	0.81
834	17	1	117	93.2	6.8	100.0	0.6	0.6	0.49	0.50	0.86	0.86
834	18	1	136	100.0	0.0	100.0	0.1	0.1	0.36	0.36	1.00	1.00
834	19	1	132	99.2	0.8	100.0	0.4	0.4	0.48	0.48	0.98	0.98

Table 5. First and Second Read Descriptive Statistics and Agreement *(continued)*

				Agreement (%)			Raw Score Mean		Raw Score Standard Deviation			
Form	Item	Score Points	Total N-Count	Exact	Adj.	Total	First Read	Second Read	First Read	Second Read	Intraclass Correla-tion	Wt. Kappa
835	12	1	131	95.4	4.6	100.0	0.6	0.6	0.50	0.50	0.91	0.91
835	13	1	134	96.3	3.7	100.0	0.4	0.4	0.50	0.50	0.92	0.92
835	14	1	134	93.3	6.7	100.0	0.6	0.6	0.49	0.49	0.86	0.86
835	15	1	123	99.2	0.8	100.0	0.4	0.4	0.50	0.50	0.98	0.98
835	16	1	123	97.6	2.4	100.0	0.5	0.5	0.50	0.50	0.95	0.95
835	17	1	123	95.9	4.1	100.0	0.6	0.6	0.49	0.49	0.91	0.91
835	18	1	123	95.9	4.1	100.0	0.4	0.4	0.48	0.48	0.91	0.91
835	19	1	123	97.6	2.4	100.0	0.5	0.5	0.50	0.50	0.95	0.95
836	12	1	130	93.1	6.9	100.0	0.4	0.4	0.49	0.50	0.86	0.86
836	13	1	130	97.7	2.3	100.0	0.8	0.8	0.43	0.42	0.94	0.94
836	14	1	137	99.3	0.7	100.0	0.6	0.6	0.50	0.50	0.99	0.99
836	15	1	132	92.4	7.6	100.0	0.7	0.7	0.45	0.45	0.81	0.81
836	16	1	132	87.9	12.1	100.0	0.8	0.8	0.37	0.37	0.56	0.56
836	17	1	130	97.7	2.3	100.0	0.4	0.4	0.49	0.49	0.95	0.95
836	18	1	130	98.5	1.5	100.0	0.7	0.7	0.47	0.47	0.96	0.96
836	19	1	132	95.5	4.5	100.0	0.7	0.7	0.46	0.47	0.89	0.89
837	12	1	131	96.9	3.1	100.0	0.8	0.8	0.43	0.41	0.92	0.91
837	13	1	131	95.4	4.6	100.0	0.3	0.3	0.46	0.47	0.89	0.89
837	14	1	131	92.4	7.6	100.0	0.8	0.8	0.43	0.42	0.79	0.79
837	15	1	120	92.5	7.5	100.0	0.8	0.8	0.43	0.42	0.79	0.79
837	16	1	120	93.3	6.7	100.0	0.3	0.3	0.47	0.46	0.85	0.85
837	17	1	114	93.0	7.0	100.0	0.5	0.5	0.50	0.50	0.86	0.86
837	18	1	114	91.2	8.8	100.0	0.7	0.7	0.46	0.44	0.78	0.78
837	19	1	120	97.5	2.5	100.0	0.5	0.5	0.50	0.50	0.95	0.95
837	20	1	120	93.3	6.7	100.0	0.6	0.5	0.50	0.50	0.87	0.87
838	12	1	126	94.4	5.6	100.0	0.6	0.6	0.48	0.48	0.88	0.88
838	13	1	126	92.1	7.9	100.0	0.5	0.5	0.50	0.50	0.84	0.84
838	14	1	131	100.0	0.0	100.0	0.4	0.4	0.50	0.50	1.00	1.00
838	16	1	128	95.3	4.7	100.0	0.3	0.3	0.48	0.47	0.90	0.89
838	17	1	128	96.9	3.1	100.0	0.4	0.4	0.50	0.50	0.94	0.94
838	18	1	128	93.8	6.2	100.0	0.3	0.3	0.48	0.48	0.86	0.86
838	19	1	128	94.5	5.5	100.0	0.4	0.4	0.49	0.49	0.89	0.89
838	20	1	128	96.9	3.1	100.0	0.4	0.4	0.49	0.48	0.93	0.93
839	12	1	122	97.5	2.5	100.0	0.6	0.6	0.48	0.49	0.95	0.95
839	13	1	122	92.6	7.4	100.0	0.8	0.7	0.43	0.45	0.82	0.81
839	14	1	122	96.7	3.3	100.0	0.3	0.3	0.44	0.46	0.92	0.92
839	15	1	122	93.4	6.6	100.0	0.7	0.7	0.47	0.46	0.85	0.85

Table 5. First and Second Read Descriptive Statistics and Agreement *(continued)*

				Agreement (%)			Raw Score Mean		Raw Score Standard Deviation			
Form	Item	Score Points	Total N-Count	Exact	Adj.	Total	First Read	Second Read	First Read	Second Read	Intraclass Correla-tion	Wt. Kappa
839	16	1	130	91.5	8.5	100.0	0.4	0.4	0.49	0.48	0.82	0.82
839	17	1	130	92.3	7.7	100.0	0.5	0.5	0.50	0.50	0.85	0.85
839	18	1	130	98.5	1.5	100.0	0.3	0.3	0.48	0.47	0.97	0.97
839	19	1	130	94.6	5.4	100.0	0.3	0.3	0.47	0.48	0.88	0.88
839	20	1	130	91.5	8.5	100.0	0.6	0.7	0.48	0.47	0.82	0.81
840	13	1	129	98.4	1.6	100.0	0.5	0.5	0.50	0.50	0.97	0.97
840	14	1	129	98.4	1.6	100.0	0.3	0.3	0.48	0.48	0.97	0.97
840	15	1	129	93.8	6.2	100.0	0.6	0.6	0.50	0.50	0.87	0.87
840	16	1	124	97.6	2.4	100.0	0.2	0.2	0.41	0.42	0.93	0.93
840	17	1	124	93.5	6.5	100.0	0.4	0.4	0.48	0.48	0.86	0.86
840	18	1	131	100.0	0.0	100.0	0.4	0.4	0.49	0.49	1.00	1.00
840	19	1	131	97.7	2.3	100.0	0.3	0.4	0.47	0.48	0.95	0.95
840	20	1	129	99.2	0.8	100.0	0.5	0.5	0.50	0.50	0.98	0.98
841	13	2	133	91.0	8.3	99.2	1.0	1.0	0.87	0.86	0.92	0.90
841	14	1	128	97.7	2.3	100.0	0.5	0.5	0.50	0.50	0.95	0.95
841	15	2	128	88.3	11.7	100.0	1.1	1.1	0.85	0.83	0.92	0.87
841	16	2	129	94.6	5.4	100.0	0.8	0.8	0.73	0.72	0.95	0.93

* Adj. = difference of 1

Constructed-Response Item Means and Standard Deviations

The average score for each constructed-response item was computed based on the first and second reads. In addition, the standard deviation of the scores was computed.

Table 5 contains the means and standard deviations for the first and second read scores. The largest difference between the item means for the first and second scores was 0.1, while there were minimal differences among standard deviation statistics.

Intraclass Correlation

The intraclass correlation was computed for each item. This correlation is an estimate of the reliability of scoring based on an average of the first and second read scores. Correlations greater than 0.60 are considered very strong because they explain more than one-third of the variance in scores. All but one item had intraclass correlations greater than or equal to 0.70 (See Table 5). Consistent with other information provided in the table, these values indicate a very high level of scoring reliability.

Weighted Kappa

Weighted Kappa (Cohen, 1968) was calculated for each item based on the first and second reads. This statistic produces an estimate of the reliability of the score classifications relative to what would be expected to occur by chance.

Weighted Kappa is an estimate of the reliability of the score classifications. That is, the Kappa statistic is a measure of reproducibility for categorical data. Guidelines for the evaluation of this statistic are:

- $k > 0.75$ denotes excellent reproducibility
- $0.4 < k \leq 0.75$ denotes good reproducibility
- $0 < k \leq 0.4$ denotes marginal reproducibility

The results found in Table 5 show a high degree of consistency between the first and second reads. For all but one item, the Weighted Kappa statistics ranged from 0.56 to 1.00, which in all cases indicates good-to-excellent reproducibility.

Based on the scoring reliability analyses, there is strong evidence that the scoring of the constructed-response items was performed in a highly reliable manner.

ITEM RESPONSE THEORY (IRT) STATISTICS

As discussed above, the item mean is a statistic used to evaluate item difficulty. However, many different test forms are used during field testing and different samples of students are responding to these items. The average ability of the different samples of students varies and a direct comparison of item means across test forms may lead to inaccurate interpretations. Therefore, Item Response Theory (IRT) was also used to evaluate item difficulty.

Specifically, the Rasch Partial Credit Model (PCM) (Masters, 1982) was used. With the use of this model, the difficulty of items and the ability of examinees are placed on the same metric. Thus, the difficulty of an item and the ability of a person can be meaningfully compared across field test forms. Also, the use of this model provides greater flexibility in situations where different samples or test forms are used, because the parameters generated are generally not considered to be sample dependent or test dependent. A description of this model, results of item calibration, and item fit evaluation are presented below.

The PCM provides an overall difficulty estimate for each item. Specifically for constructed-response items when there are several points possible, individual estimates of difficulty for each of the possible score points are also calculated (i.e., step values). Each step value represents the difficulty of a student receiving a particular score point, given that he or she has already received the prior score point. For example, if a 3-point item had step values of −1.0, 1.0, and 0.0, one could say that it is relatively easy to obtain a score of 1. However, it is much more difficult to obtain a 2, given the student

has the ability to score a 1, because the difference in difficulty between a 1 and a 2 is much greater than the difference between a 0 and a 1. Also, the difference between a 2 and a 3 is not as great as the difference between a 1 and a 2. Thus, with this example, a small step is needed to go from a 0 to a 1, a large step is needed to move from a 1 to a 2, and a moderate step is needed to proceed from a 2 to a 3.

Item Calibration

As discussed above, the use of Rasch item difficulty statistics provides an advantage over the use of classical item means because they can be compared across test forms. Students from different samples responded to the various test forms. Although the samples were selected to be similar with respect to student ability, there are differences. By equating the test forms (See the Equating Procedure section below), the Rasch item difficulties account for those differences, and these statistics can be compared across test forms.

Rasch item difficulty values generally range from −3.00 to +3.00. An item with a Rasch difficulty greater than +2.0 is considered very difficult and should be examined carefully. If the item is measuring an important concept that students are having difficulty with, then the item can be useful. However, if the item is measuring a trivial concept or is written in a confusing manner, then it might not be appropriate to use on an operational test form. Likewise, any item with a Rasch difficulty less than −2.0 is considered very easy and usually provides little information regarding student achievement. The vast majority of test items should range between −2.0 and +2.0. This range represents approximately two standard deviations around the average difficulty of 0. Thus, one would expect that, based on chance, roughly 5% of the items will fall outside of that range and, therefore, these are items that should be closely examined for content.

Item Fit Evaluation

The INFIT statistic is used to determine whether items are functioning in a way that is congruent with the assumptions of the Rasch model. Under these assumptions, how a student will respond to an item depends on the proficiency of the student and the difficulty of the item, both of which are on the same measurement scale. If an item is as difficult as a student is able, the student will have a 50% chance of getting the item correct. If a student is more able than an item is difficult, under the assumptions of the Rasch model, that student has a greater than 50% chance of correctly answering the item. On the other hand, if the item is more difficult than the student is able, he or she has a less than 50% chance of correctly responding to the item. Rasch fit statistics estimate the extent to which an item is functioning in this predicted manner. Items showing a poor fit with the Rasch model typically have values outside the range of 0.7 to 1.3.

Table 6 contains a summary of the Partial Credit Model item analysis for each of the field test forms. The first column lists the form numbers. The next two columns list the

number of students who participated and the number of items on each field test form, respectively. The remaining columns are divided into two sections. The first section pertains to the Rasch item difficulties, while the second pertains to the INFIT statistics. Most of the items fell within the moderate −2.0 to +2.0 difficulty range, and only four items had an INFIT statistic outside the typical range.

Table 6. Partial Credit Model Item Analysis

			Rasch			INFIT		
Form	N-Count	No. of Items	<−2.0	−2.0 to 2.0	>2.0	<−0.70	−0.70 to 1.30	>1.30
831	775	20	0	19	1	0	20	0
832	770	19	0	19	0	0	19	0
833	791	19	2	17	0	0	18	1
834	774	19	1	17	1	0	19	0
835	778	19	1	18	0	0	18	1
836	767	19	0	19	0	0	19	0
837	752	20	0	20	0	0	19	1
838	766	20	0	19	0	0	18	1
839	757	20	0	20	0	0	20	0
840	756	20	0	20	0	0	20	0
841	770	16	1	15	0	0	16	0

* For some forms, the item counts in the “Rasch” and “INFIT” columns may not sum to the value in the “No. of Items” column due to DNS (do not score) items.

All of the individual IRT item statistics are provided in Appendix B. The column entitled “RID” contains the Rasch item difficulty statistics. S1–S6 contain the step values for the constructed-response items. Finally, INFIT contains the INFIT statistic for each item.

DIFFERENTIAL ITEM FUNCTIONING (DIF) STATISTICS

Statistical procedures are employed to observe whether, on the basis of data, there exists the possibility of unfair treatment of different populations. DIF statistics are used to identify items for which members of a focal group have a different probability of getting the items correct than members of a reference group after the groups have been matched on ability level on the test.

For the multiple-choice items, the Mantel-Haenszel Delta (MHD) DIF statistics were computed (Dorans & Holland, 1992) to classify test items in three levels of DIF for each comparison: negligible DIF (A), moderate DIF (B), and large DIF (C). An item was flagged if it exhibited a B or C category of DIF, using the following rules derived from

National Assessment of Educational Progress (NAEP) guidelines (Allen, Carlson, & Zalanak, 1999):

- MHD not significantly different from 0 (based on alpha = 0.05) **or** |MHD| < 1.0 are classified as A.

- MHD significantly different from 0 and {|MHD| ≥ 1.0 and < 1.5} **or**
  MHD not significantly different from 0 and |MHD| ≥ 1.0 are classified as B.

- |MHD| ≥ 1.5 and significantly different from 0 are classified as C.

For the constructed-response items, the effect size of the standardized mean difference (SMD) was used to flag DIF. The SMD reflects the size of the differences in performance on constructed-response items between student groups matched on the total score. It is the difference between the unweighted item mean of the focal group and the weighted item mean of the reference group. The weights applied to the reference group are applied so that the weighted number of reference group students is the same as in the focal group (within the same ability group). The SMD is divided by the total group item standard deviation to get a measure of the effect size (ES) for the SMD. The SMD effect size groups each item into one of three categories: negligible DIF (AA), moderate DIF (BB), and large DIF (CC). Only categories BB and CC were flagged in the results.

- Probability is > 0.05 **or** if |ES| is ≤ 0.17, classified as AA.

- Probability is > 0.05 and if 0.17 < |ES| ≤ 0.25, classified as BB.

- Probability is > 0.05 and if |ES| is > 0.25, classified as CC.

Although DIF statistics are typically conducted by gender and ethnicity, the low n-counts for ethnic subgroups did not allow for these statistics to be meaningful. The n-counts for gender allowed for comparisons to be made, but were still somewhat low, so resulting statistics should be interpreted with caution.

The DIF statistics for gender are shown in Appendix C. Flagging of items appears in the “DIF Category” column, and if an item is flagged, the “Favored Group” column indicates which gender is favored.

Section III: Equating Procedure

The 2011 field test administration for the New York State Regents Examination in Physics consisted of 10 field test forms numbered 831–840 and one anchor form labeled 841. The field test forms contained multiple-choice and constructed-response items. Each student participating in the field test was administered one of the 11 test forms. The test forms were spiraled within the classroom so that the groups of students taking each form were equivalent. A complete listing of these field test forms can be seen in Appendix A, where item type (e.g., multiple-choice, constructed-response) and the maximum points for each item are displayed.

The anchor form was equated to the item bank using a common-item equating design. The anchor-item difficulty parameters were fixed to their 2010 item bank values. This places the item difficulty estimates and the ability estimates of the students taking the anchor form onto the item bank scale. After the anchor form was placed onto the bank scale, the average of the two mean ability estimates for the two forms was computed using ability estimates of nonextreme students. This average ability estimate was used to equate the remaining field test forms, as well as to update the item parameters for the anchor form.

As a part of the anchor item equating, an item-stability check was performed. After fixing all of the items to their 2010 bank values, any item with a displacement value with a magnitude greater than 0.30 was no longer fixed, and the test form was reanalyzed. If more than one item had a displacement value with a magnitude greater than 0.30, then the item with the largest displacement was freed, and the test form was reanalyzed. In a stepwise fashion, this procedure was repeated until all remaining fixed anchor items had displacements with magnitudes less than or equal to 0.30.

Applying the anchor item-stability check to the anchor form resulted in no items having a displacement value with a magnitude greater than 0.30. This indicates a strong level of stability in the items used on the anchor form.

The equated mean ability estimate for form 841 was −0.10. This value served as the target mean ability for the equating process.

After the anchor form was equated and the target mean was computed, the field test forms were equated using the equivalent groups design. The first step was to calibrate each form separately where all the item parameters were free to estimate (without constraint). From those initial calibrations, the mean ability estimates for each field test form were obtained. The second step was to determine the equating constant for each form by subtracting the mean ability for a given field test form from the target mean ability calculated from the anchor form (i.e., form 841). The respective equating constant was then added to each of the item parameters on a given form. If the resulting mean of the ability estimates for those students did not equal that of the target mean, then the procedure was repeated until the mean abilities for each of the field test forms equaled

the target mean ability. Table 7 shows the mean abilities and constants used for the equating.

Table 7. Initial Mean Abilities and Equating Constants

Form Number	Mean Ability	Constant
831	0.04	−0.14
832	0.22	−0.30
833	0.32	−0.40
834	0.18	−0.27
835	0.22	−0.31
836	0.30	−0.38
837	0.50	−0.57
838	−0.26	0.15
839	0.26	−0.35
840	−0.05	−0.05

The equated item parameters for the field test items can now be compared across test forms, since the equating process places all items on the same scale. In addition, when items are combined to form unique operational test forms, raw score-to-scale score tables can be generated based on these parameters. The following section contains a description of the development of the operational test forms and scoring tables.

Section IV: Scaling of Operational Test Forms

Operational test items are selected based on content coverage and statistical quality. The sets of items on each operational test conform to the coverage suggested by content experts. These expert judgments are based on the learning standards established by the New York State Education Department. With respect to statistical quality, classical and Rasch statistics are examined to determine how well items function. Also, items are selected such that they range in difficulty in order to measure students across ability levels. Appendix D contains the 2011 operational test maps with content information regarding each item included on the January 2011 and June 2011 operational test forms.

In order to limit wide fluctuations of raw scores that correspond to scale scores of 65 and 85 across administrations, the average Rasch item difficulty for the operational test is considered. For this examination, an average Rasch difficulty of approximately −0.319 is used as a target for each administration. In most cases, meeting this target will provide raw scores of similar magnitude to other forms. However, differences with these scores also occur due to the distribution of the Rasch item difficulty parameters.

Scoring tables display the relationship between raw scores on the operational test and assigned scale scores. Appendix E contains the scoring tables used for the January and June 2011 operational test forms. Four steps are taken in order to produce these tables and the resulting conversion charts.

The first step is to develop a raw score (i.e., number of points on the test form) to theta (i.e., student ability) to scale score relationship for the baseline operational test form. This relationship is determined when standards are set and then used for every administration moving forward until the standards are revisited. The baseline form was determined by the New York State Education Department to be June 2004. The raw score-to-theta relationship from that examination was used, and then scale scores are calculated based on the raw score cuts according to the following formula:

$$p(x) = m_3x^3 + m_2x^2 + m_1x + m_0$$

The raw score of zero was assigned a scale score of zero, and the maximum raw score was assigned a scale score of 100. The raw scores corresponding to the scale scores of 65 and 85 were also fixed. The polynomial relationship shown above was then used to assign all scale scores to the remaining raw scores. The resulting values for m_1–m_3 are the transformation constants used to produce the final raw score-to-scale score table.

The second step is to develop a raw score-to-theta relationship for the new operational test form, using the field test equated PCM item parameters. This is accomplished by doing a calibration where all items are anchored to their field test parameters. The number of points on the test form (i.e., raw score) expected across student ability levels is based on the difficulty of the items on the form. Thus, given a particular student ability level (i.e., theta), if the points are more difficult to earn on the new test than the points on the June 2004 test, the number of points expected of this student on the new test will be less than the number of points, expected of this student on the baseline form.

The third step is to use linear interpolation to determine the raw score-to-theta-to-scale score relationship for the new test. The theta values associated with scale scores of 65 and 85 on the baseline form are used along with the raw score-to-theta relationship developed in the previous step. In other words, the baseline 65 and 85 theta values are used as reference points, and linear interpolation assigns the other scale scores.

Finally, a conversion chart is created based on the scoring table generated in the third step. Scale scores are rounded to the nearest whole number in all cases except for 0, 65, 85, and 100. A raw score of zero is assigned a scale score of zero. The maximum raw score is assigned a scale score of 100. With respect to the 65 and 85 scale scores, the raw scores with scale scores of 65 or 85, after rounding, are assigned those values.

References

Allen, N. L., Carlson, J. E., and Zalanak, C. A. (1999). *The NAEP 1996 technical report.* Washington, DC: National Center for Education Statistics.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning: Theory and practice (*35–66). Hillsdale, NJ: Erlbaum.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Appendix A: Classical Item Analysis

Table 8. Classical Item Analysis

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2011_Phys_FT	831	MC	01	1	775	0.78	0.00		0.29	0.07	0.14	0.49			0.49	0.42
2011_Phys_FT	831	MC	02	1	775	0.78	0.03		0.32	0.35	0.22	0.08			0.32	0.44
2011_Phys_FT	831	MC	03	1	775	0.78	0.00		0.05	0.12	0.33	0.50			0.33	0.23
2011_Phys_FT	831	MC	04	1	775	0.78	0.01		0.32	0.10	0.05	0.52			0.52	0.39
2011_Phys_FT	831	MC	05	1	775	0.78	0.01		0.43	0.35	0.12	0.09			0.35	0.27
2011_Phys_FT	831	MC	06	1	775	0.78	0.00		0.08	0.06	0.09	0.77			0.77	0.44
2011_Phys_FT	831	MC	07	1	775	0.78	0.02		0.14	0.16	0.64	0.05			0.64	0.36
2011_Phys_FT	831	MC	08	1	775	0.78	0.02		0.58	0.25	0.10	0.05			0.58	0.49
2011_Phys_FT	831	MC	09	1	775	0.78	0.01		0.09	0.10	0.18	0.62			0.62	0.35
2011_Phys_FT	831	MC	10	1	775	0.78	0.01		0.73	0.06	0.03	0.18			0.73	0.39
2011_Phys_FT	831	MC	11	1	775	0.78	0.00		0.30	0.07	0.56	0.07			0.56	0.36
2011_Phys_FT	831	MC	12	1	775	0.78	0.02		0.13	0.48	0.30	0.08			0.48	0.43
2011_Phys_FT	831	CR	13	1	775	0.78	0.03	0.37	0.60						0.60	0.50
2011_Phys_FT	831	CR	14	1	775	0.78	0.07	0.44	0.49						0.49	0.38
2011_Phys_FT	831	CR	15	1	775	0.78	0.07	0.29	0.63						0.63	0.50
2011_Phys_FT	831	CR	16	1	775	0.78	0.11	0.25	0.65						0.65	0.55
2011_Phys_FT	831	CR	17	1	775	0.78	0.18	0.37	0.46						0.46	0.61
2011_Phys_FT	831	CR	18	1	775	0.78	0.18	0.42	0.39						0.39	0.65
2011_Phys_FT	831	CR	19	1	775	0.78	0.22	0.67	0.10						0.10	0.44
2011_Phys_FT	831	CR	20	1	775	0.78	0.23	0.30	0.47						0.47	0.62
2011_Phys_FT	832	MC	01	1	770	0.76	0.01		0.47	0.24	0.17	0.11			0.47	0.44
2011_Phys_FT	832	MC	02	1	770	0.76	0.00		0.02	0.06	0.31	0.61			0.61	0.31

Table 8. Classical Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2011_Phys_FT	832	MC	03	1	770	0.76	0.00		0.49	0.24	0.23	0.03			0.49	0.39
2011_Phys_FT	832	MC	04	1	770	0.76	0.00		0.33	0.02	0.61	0.04			0.61	0.33
2011_Phys_FT	832	MC	05	1	770	0.76	0.00		0.03	0.13	0.05	0.78			0.78	0.39
2011_Phys_FT	832	MC	06	1	770	0.76	0.00		0.28	0.62	0.06	0.04			0.62	0.22
2011_Phys_FT	832	MC	07	1	770	0.76	0.00		0.15	0.13	0.42	0.30			0.42	0.31
2011_Phys_FT	832	MC	08	1	770	0.76	0.01		0.04	0.65	0.15	0.14			0.65	0.40
2011_Phys_FT	832	MC	09	1	770	0.76	0.01		0.09	0.13	0.08	0.69			0.69	0.52
2011_Phys_FT	832	MC	10	1	770	0.76	0.01		0.33	0.05	0.55	0.06			0.55	0.46
2011_Phys_FT	832	MC	11	1	770	0.76	0.04		0.41	0.24	0.21	0.10			0.41	0.43
2011_Phys_FT	832	CR	12	1	770	0.76	0.03	0.36	0.61						0.61	0.50
2011_Phys_FT	832	CR	13	1	770	0.76	0.04	0.49	0.47						0.47	0.45
2011_Phys_FT	832	CR	14	1	770	0.76	0.06	0.27	0.67						0.67	0.51
2011_Phys_FT	832	CR	15	1	770	0.76	0.06	0.30	0.64						0.64	0.48
2011_Phys_FT	832	CR	16	1	770	0.76	0.12	0.71	0.16						0.16	0.43
2011_Phys_FT	832	CR	17	1	770	0.76	0.21	0.31	0.48						0.48	0.61
2011_Phys_FT	832	CR	18	1	770	0.76	0.21	0.29	0.50						0.50	0.62
2011_Phys_FT	832	CR	19	1	770	0.76	0.10	0.39	0.51						0.51	0.41
2011_Phys_FT	833	MC	01	1	791	0.71	0.00		0.49	0.47	0.02	0.02			0.49	0.42
2011_Phys_FT	833	MC	02	1	791	0.71	0.00		0.12	0.04	0.17	0.67			0.67	0.30
2011_Phys_FT	833	MC	03	1	791	0.71	0.01		0.05	0.41	0.24	0.30			0.41	0.31
2011_Phys_FT	833	MC	04	1	791	0.71	0.00		0.04	0.06	0.62	0.28			0.62	0.39
2011_Phys_FT	833	MC	05	1	791	0.71	0.01		0.54	0.08	0.34	0.03			0.54	0.24
2011_Phys_FT	833	MC	06	1	791	0.71	0.00		0.01	0.20	0.72	0.07			0.72	0.33

Table 8. Classical Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2011_Phys_FT	833	MC	07	1	791	0.71	0.00		0.22	0.13	0.49	0.16			0.49	0.43
2011_Phys_FT	833	MC	08	1	791	0.71	0.05		0.15	0.46	0.25	0.10			0.46	0.34
2011_Phys_FT	833	MC	09	1	791	0.71	0.02		0.33	0.26	0.22	0.18			0.26	0.04
2011_Phys_FT	833	MC	10	1	791	0.71	0.00		0.04	0.90	0.04	0.02			0.90	0.24
2011_Phys_FT	833	MC	11	1	791	0.71	0.01		0.05	0.03	0.70	0.22			0.70	0.32
2011_Phys_FT	833	CR	12	1	791	0.71	0.02	0.13	0.85						0.85	0.38
2011_Phys_FT	833	CR	13	1	791	0.71	0.15	0.51	0.34						0.34	0.53
2011_Phys_FT	833	CR	14	1	791	0.71	0.16	0.22	0.61						0.61	0.56
2011_Phys_FT	833	CR	15	1	791	0.71	0.07	0.23	0.70						0.70	0.48
2011_Phys_FT	833	CR	16	1	791	0.71	0.08	0.50	0.43						0.43	0.56
2011_Phys_FT	833	CR	17	1	791	0.71	0.22	0.48	0.30						0.30	0.52
2011_Phys_FT	833	CR	18	1	791	0.71	0.22	0.25	0.53						0.53	0.62
2011_Phys_FT	833	CR	19	1	791	0.71	0.15	0.34	0.50						0.50	0.53
2011_Phys_FT	834	MC	01	1	774	0.77	0.01		0.08	0.21	0.59	0.11			0.59	0.43
2011_Phys_FT	834	MC	02	1	774	0.77	0.02		0.09	0.17	0.64	0.07			0.64	0.44
2011_Phys_FT	834	MC	03	1	774	0.77	0.01		0.04	0.85	0.05	0.05			0.85	0.33
2011_Phys_FT	834	MC	04	1	774	0.77	0.00		0.07	0.16	0.05	0.71			0.71	0.21
2011_Phys_FT	834	MC	05	1	774	0.77	0.02		0.11	0.11	0.15	0.61			0.61	0.51
2011_Phys_FT	834	MC	06	1	774	0.77	0.01		0.47	0.22	0.13	0.17			0.47	0.42
2011_Phys_FT	834	MC	07	1	774	0.77	0.01		0.25	0.53	0.07	0.14			0.53	0.49
2011_Phys_FT	834	MC	08	1	774	0.77	0.01		0.25	0.60	0.07	0.06			0.60	0.46
2011_Phys_FT	834	MC	09	1	774	0.77	0.03		0.28	0.37	0.18	0.13			0.28	0.40
2011_Phys_FT	834	MC	10	1	774	0.77	0.01		0.05	0.05	0.21	0.68			0.68	0.45

Table 8. Classical Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2011_Phys_FT	834	MC	11	1	774	0.77	0.02		0.05	0.51	0.08	0.35			0.51	0.38
2011_Phys_FT	834	CR	12	1	774	0.77	0.04	0.26	0.70						0.70	0.49
2011_Phys_FT	834	CR	13	1	774	0.77	0.06	0.39	0.55						0.55	0.52
2011_Phys_FT	834	CR	14	1	774	0.77	0.07	0.38	0.54						0.54	0.55
2011_Phys_FT	834	CR	15	1	774	0.77	0.13	0.19	0.68						0.68	0.40
2011_Phys_FT	834	CR	16	1	774	0.77	0.22	0.52	0.25						0.25	0.57
2011_Phys_FT	834	CR	17	1	774	0.77	0.22	0.26	0.52						0.52	0.63
2011_Phys_FT	834	CR	18	1	774	0.77	0.13	0.74	0.13						0.13	0.36
2011_Phys_FT	834	CR	19	1	774	0.77	0.15	0.49	0.36						0.36	0.31
2011_Phys_FT	835	MC	01	1	778	0.80	0.00		0.21	0.07	0.66	0.05			0.66	0.40
2011_Phys_FT	835	MC	02	1	778	0.80	0.00		0.02	0.08	0.02	0.87			0.87	0.30
2011_Phys_FT	835	MC	03	1	778	0.80	0.00		0.09	0.04	0.35	0.52			0.52	0.35
2011_Phys_FT	835	MC	04	1	778	0.80	0.00		0.67	0.10	0.19	0.04			0.67	0.43
2011_Phys_FT	835	MC	05	1	778	0.80	0.01		0.05	0.37	0.51	0.06			0.51	0.35
2011_Phys_FT	835	MC	06	1	778	0.80	0.00		0.01	0.05	0.27	0.67			0.67	0.29
2011_Phys_FT	835	MC	07	1	778	0.80	0.01		0.05	0.60	0.07	0.27			0.60	0.48
2011_Phys_FT	835	MC	08	1	778	0.80	0.01		0.23	0.11	0.35	0.30			0.30	0.39
2011_Phys_FT	835	MC	09	1	778	0.80	0.01		0.04	0.69	0.07	0.20			0.69	0.44
2011_Phys_FT	835	MC	10	1	778	0.80	0.01		0.50	0.41	0.04	0.03			0.50	0.31
2011_Phys_FT	835	MC	11	1	778	0.80	0.03		0.35	0.33	0.18	0.11			0.35	0.20
2011_Phys_FT	835	CR	12	1	778	0.80	0.03	0.44	0.53						0.53	0.48
2011_Phys_FT	835	CR	13	1	778	0.80	0.07	0.47	0.46						0.46	0.51
2011_Phys_FT	835	CR	14	1	778	0.80	0.07	0.29	0.63						0.63	0.52

Table 8. Classical Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2011_Phys_FT	835	CR	15	1	778	0.80	0.17	0.42	0.42						0.42	0.58
2011_Phys_FT	835	CR	16	1	778	0.80	0.13	0.37	0.50						0.50	0.68
2011_Phys_FT	835	CR	17	1	778	0.80	0.14	0.27	0.59						0.59	0.70
2011_Phys_FT	835	CR	18	1	778	0.80	0.22	0.46	0.32						0.32	0.69
2011_Phys_FT	835	CR	19	1	778	0.80	0.22	0.29	0.48						0.48	0.65
2011_Phys_FT	836	MC	01	1	767	0.75	0.01		0.21	0.16	0.23	0.39			0.39	0.45
2011_Phys_FT	836	MC	02	1	767	0.75	0.00		0.46	0.02	0.02	0.51			0.51	0.41
2011_Phys_FT	836	MC	03	1	767	0.75	0.01		0.17	0.09	0.66	0.08			0.66	0.44
2011_Phys_FT	836	MC	04	1	767	0.75	0.02		0.05	0.53	0.24	0.16			0.53	0.38
2011_Phys_FT	836	MC	05	1	767	0.75	0.02		0.12	0.48	0.14	0.23			0.48	0.36
2011_Phys_FT	836	MC	06	1	767	0.75	0.01		0.12	0.65	0.15	0.07			0.65	0.47
2011_Phys_FT	836	MC	07	1	767	0.75	0.01		0.38	0.47	0.07	0.06			0.47	0.36
2011_Phys_FT	836	MC	08	1	767	0.75	0.01		0.10	0.37	0.20	0.33			0.37	0.34
2011_Phys_FT	836	MC	09	1	767	0.75	0.02		0.17	0.13	0.48	0.20			0.48	0.35
2011_Phys_FT	836	MC	10	1	767	0.75	0.03		0.28	0.43	0.20	0.06			0.43	0.34
2011_Phys_FT	836	MC	11	1	767	0.75	0.03		0.20	0.15	0.08	0.55			0.55	0.39
2011_Phys_FT	836	CR	12	1	767	0.75	0.09	0.46	0.45						0.45	0.55
2011_Phys_FT	836	CR	13	1	767	0.75	0.10	0.15	0.76						0.76	0.55
2011_Phys_FT	836	CR	14	1	767	0.75	0.04	0.38	0.58						0.58	0.45
2011_Phys_FT	836	CR	15	1	767	0.75	0.06	0.18	0.77						0.77	0.35
2011_Phys_FT	836	CR	16	1	767	0.75	0.06	0.13	0.81						0.81	0.37
2011_Phys_FT	836	CR	17	1	767	0.75	0.14	0.47	0.39						0.39	0.51
2011_Phys_FT	836	CR	18	1	767	0.75	0.14	0.22	0.64						0.64	0.60

Table 8. Classical Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2011_Phys_FT	836	CR	19	1	767	0.75	0.14	0.19	0.68						0.68	0.46
2011_Phys_FT	837	MC	01	1	752	0.77	0.01		0.68	0.05	0.16	0.10			0.68	0.35
2011_Phys_FT	837	MC	02	1	752	0.77	0.01		0.06	0.06	0.09	0.79			0.79	0.47
2011_Phys_FT	837	MC	03	1	752	0.77	0.04		0.43	0.16	0.22	0.15			0.43	0.43
2011_Phys_FT	837	MC	04	1	752	0.77	0.00		0.04	0.20	0.06	0.70			0.70	0.38
2011_Phys_FT	837	MC	05	1	752	0.77	0.00		0.61	0.33	0.05	0.01			0.61	0.26
2011_Phys_FT	837	MC	06	1	752	0.77	0.01		0.03	0.75	0.13	0.07			0.75	0.46
2011_Phys_FT	837	MC	07	1	752	0.77	0.00		0.02	0.11	0.13	0.73			0.73	0.36
2011_Phys_FT	837	MC	08	1	752	0.77	0.01		0.41	0.10	0.33	0.14			0.41	0.15
2011_Phys_FT	837	MC	09	1	752	0.77	0.00		0.05	0.07	0.69	0.19			0.69	0.32
2011_Phys_FT	837	MC	10	1	752	0.77	0.03		0.35	0.10	0.44	0.09			0.35	0.34
2011_Phys_FT	837	MC	11	1	752	0.77	0.01		0.80	0.12	0.04	0.03			0.80	0.47
2011_Phys_FT	837	CR	12	1	752	0.77	0.03	0.14	0.83						0.83	0.50
2011_Phys_FT	837	CR	13	1	752	0.77	0.06	0.66	0.28						0.28	0.47
2011_Phys_FT	837	CR	14	1	752	0.77	0.06	0.25	0.70						0.70	0.54
2011_Phys_FT	837	CR	15	1	752	0.77	0.09	0.19	0.72						0.72	0.52
2011_Phys_FT	837	CR	16	1	752	0.77	0.11	0.61	0.29						0.29	0.43
2011_Phys_FT	837	CR	17	1	752	0.77	0.16	0.40	0.45						0.45	0.51
2011_Phys_FT	837	CR	18	1	752	0.77	0.16	0.22	0.63						0.63	0.56
2011_Phys_FT	837	CR	19	1	752	0.77	0.15	0.33	0.51						0.51	0.55
2011_Phys_FT	837	CR	20	1	752	0.77	0.18	0.30	0.52						0.52	0.52
2011_Phys_FT	838	MC	01	1	766	0.73	0.01		0.20	0.61	0.08	0.09			0.61	0.44
2011_Phys_FT	838	MC	02	1	766	0.73	0.00		0.32	0.02	0.37	0.29			0.32	0.41

Table 8. Classical Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2011_Phys_FT	838	MC	03	1	766	0.73	0.01		0.11	0.16	0.38	0.35			0.35	0.29
2011_Phys_FT	838	MC	04	1	766	0.73	0.03		0.09	0.43	0.33	0.13			0.33	0.09
2011_Phys_FT	838	MC	05	1	766	0.73	0.01		0.14	0.17	0.59	0.09			0.59	0.40
2011_Phys_FT	838	MC	06	1	766	0.73	0.01		0.47	0.14	0.04	0.34			0.47	0.36
2011_Phys_FT	838	MC	07	1	766	0.73	0.01		0.16	0.50	0.18	0.17			0.50	0.34
2011_Phys_FT	838	MC	08	1	766	0.73	0.03		0.19	0.42	0.27	0.09			0.42	0.37
2011_Phys_FT	838	MC	09	1	766	0.73	0.01		0.28	0.11	0.05	0.55			0.55	0.37
2011_Phys_FT	838	MC	10	1	766	0.73	0.01		0.45	0.50	0.02	0.01			0.50	0.41
2011_Phys_FT	838	MC	11	1	766	0.73	0.03		0.27	0.39	0.08	0.24			0.39	0.30
2011_Phys_FT	838	CR	12	1	766	0.73	0.04	0.32	0.65						0.65	0.43
2011_Phys_FT	838	CR	13	1	766	0.73	0.04	0.41	0.56						0.56	0.42
2011_Phys_FT	838	CR	14	1	766	0.73	0.11	0.44	0.45						0.45	0.45
2011_Phys_FT	838	CR	15													
2011_Phys_FT	838	CR	16	1	766	0.73	0.14	0.56	0.30						0.30	0.52
2011_Phys_FT	838	CR	17	1	766	0.73	0.15	0.42	0.43						0.43	0.56
2011_Phys_FT	838	CR	18	1	766	0.73	0.29	0.37	0.33						0.33	0.59
2011_Phys_FT	838	CR	19	1	766	0.73	0.29	0.34	0.37						0.37	0.61
2011_Phys_FT	838	CR	20	1	766	0.73	0.19	0.47	0.34						0.34	0.49
2011_Phys_FT	839	MC	01	1	757	0.79	0.00		0.28	0.06	0.05	0.60			0.60	0.37
2011_Phys_FT	839	MC	02	1	757	0.79	0.02		0.17	0.08	0.60	0.12			0.60	0.24
2011_Phys_FT	839	MC	03	1	757	0.79	0.00		0.04	0.20	0.05	0.71			0.71	0.42
2011_Phys_FT	839	MC	04	1	757	0.79	0.01		0.11	0.78	0.05	0.05			0.78	0.38
2011_Phys_FT	839	MC	05	1	757	0.79	0.01		0.24	0.09	0.11	0.56			0.56	0.30

Table 8. Classical Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2011_Phys_FT	839	MC	06	1	757	0.79	0.01		0.10	0.08	0.05	0.76			0.76	0.44
2011_Phys_FT	839	MC	07	1	757	0.79	0.01		0.04	0.06	0.82	0.08			0.82	0.31
2011_Phys_FT	839	MC	08	1	757	0.79	0.01		0.05	0.27	0.58	0.08			0.58	0.35
2011_Phys_FT	839	MC	09	1	757	0.79	0.01		0.05	0.06	0.73	0.16			0.73	0.37
2011_Phys_FT	839	MC	10	1	757	0.79	0.02		0.23	0.46	0.22	0.07			0.23	0.34
2011_Phys_FT	839	MC	11	1	757	0.79	0.01		0.36	0.43	0.08	0.12			0.43	0.37
2011_Phys_FT	839	CR	12	1	757	0.79	0.05	0.41	0.53						0.53	0.60
2011_Phys_FT	839	CR	13	1	757	0.79	0.06	0.29	0.65						0.65	0.58
2011_Phys_FT	839	CR	14	1	757	0.79	0.09	0.66	0.25						0.25	0.42
2011_Phys_FT	839	CR	15	1	757	0.79	0.09	0.27	0.65						0.65	0.52
2011_Phys_FT	839	CR	16	1	757	0.79	0.14	0.47	0.39						0.39	0.56
2011_Phys_FT	839	CR	17	1	757	0.79	0.14	0.35	0.51						0.51	0.61
2011_Phys_FT	839	CR	18	1	757	0.79	0.19	0.50	0.31						0.31	0.53
2011_Phys_FT	839	CR	19	1	757	0.79	0.20	0.44	0.36						0.36	0.59
2011_Phys_FT	839	CR	20	1	757	0.79	0.20	0.19	0.61						0.61	0.61
2011_Phys_FT	840	MC	01	1	756	0.80	0.01		0.24	0.06	0.64	0.05			0.64	0.47
2011_Phys_FT	840	MC	02	1	756	0.80	0.01		0.77	0.07	0.03	0.13			0.77	0.37
2011_Phys_FT	840	MC	03	1	756	0.80	0.03		0.44	0.21	0.13	0.19			0.44	0.56
2011_Phys_FT	840	MC	04	1	756	0.80	0.01		0.44	0.20	0.29	0.05			0.29	0.44
2011_Phys_FT	840	MC	05	1	756	0.80	0.01		0.41	0.30	0.16	0.11			0.41	0.36
2011_Phys_FT	840	MC	06	1	756	0.80	0.00		0.83	0.05	0.07	0.05			0.83	0.39
2011_Phys_FT	840	MC	07	1	756	0.80	0.01		0.71	0.09	0.04	0.14			0.71	0.38
2011_Phys_FT	840	MC	08	1	756	0.80	0.02		0.50	0.06	0.34	0.08			0.50	0.36

Table 8. Classical Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2011_Phys_FT	840	MC	09	1	756	0.80	0.01		0.72	0.06	0.19	0.02			0.72	0.48
2011_Phys_FT	840	MC	10	1	756	0.80	0.01		0.05	0.08	0.63	0.23			0.63	0.32
2011_Phys_FT	840	MC	11	1	756	0.80	0.01		0.27	0.12	0.37	0.23			0.37	0.28
2011_Phys_FT	840	MC	12	1	756	0.80	0.02		0.42	0.07	0.29	0.20			0.29	0.36
2011_Phys_FT	840	CR	13	1	756	0.80	0.07	0.47	0.46						0.46	0.55
2011_Phys_FT	840	CR	14	1	756	0.80	0.11	0.59	0.30						0.30	0.57
2011_Phys_FT	840	CR	15	1	756	0.80	0.12	0.36	0.52						0.52	0.53
2011_Phys_FT	840	CR	16	1	756	0.80	0.15	0.60	0.25						0.25	0.47
2011_Phys_FT	840	CR	17	1	756	0.80	0.16	0.48	0.35						0.35	0.53
2011_Phys_FT	840	CR	18	1	756	0.80	0.28	0.36	0.36						0.36	0.60
2011_Phys_FT	840	CR	19	1	756	0.80	0.31	0.39	0.30						0.30	0.61
2011_Phys_FT	840	CR	20	1	756	0.80	0.17	0.29	0.54						0.54	0.51
2011_Phys_FT	841	MC	01	1	770	0.71	0.00		0.03	0.46	0.51	0.01			0.51	0.18
2011_Phys_FT	841	MC	02	1	770	0.71	0.01		0.06	0.08	0.18	0.66			0.66	0.51
2011_Phys_FT	841	MC	03	1	770	0.71	0.01		0.47	0.12	0.33	0.06			0.47	0.25
2011_Phys_FT	841	MC	04	1	770	0.71	0.00		0.38	0.16	0.41	0.05			0.38	0.51
2011_Phys_FT	841	MC	05	1	770	0.71	0.00		0.03	0.06	0.02	0.89			0.89	0.27
2011_Phys_FT	841	MC	06	1	770	0.71	0.00		0.06	0.24	0.14	0.56			0.56	0.37
2011_Phys_FT	841	MC	07	1	770	0.71	0.01		0.17	0.49	0.32	0.01			0.49	0.28
2011_Phys_FT	841	MC	08	1	770	0.71	0.00		0.66	0.07	0.11	0.15			0.66	0.37
2011_Phys_FT	841	MC	09	1	770	0.71	0.00		0.04	0.03	0.08	0.86			0.86	0.30
2011_Phys_FT	841	MC	10	1	770	0.71	0.00		0.01	0.04	0.43	0.51			0.51	0.29
2011_Phys_FT	841	MC	11	1	770	0.71	0.00		0.35	0.27	0.24	0.14			0.35	0.38

Table 8. Classical Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	Alpha	B	M0	M1	M2	M3	M4	M5	M6	Mean	Point-Biserial
2011_Phys_FT	841	MC	12	1	770	0.71	0.01		0.21	0.16	0.12	0.50			0.50	0.43
2011_Phys_FT	841	CR	13	2	770	0.71	0.08	0.30	0.23	0.39					1.01	0.65
2011_Phys_FT	841	CR	14	1	770	0.71	0.09	0.38	0.53						0.53	0.62
2011_Phys_FT	841	CR	15	2	770	0.71	0.09	0.18	0.28	0.44					1.17	0.68
2011_Phys_FT	841	CR	16	2	770	0.71	0.11	0.24	0.50	0.15					0.80	0.60

Appendix B: Partial Credit Model Item Analysis

Table 9. Partial Credit Model Item Analysis

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2011_Phys_FT	831	MC	01	1	775	−0.0859							1.04
2011_Phys_FT	831	MC	02	1	775	0.7823							0.99
2011_Phys_FT	831	MC	03	1	775	0.7328							1.26
2011_Phys_FT	831	MC	04	1	775	−0.2163							1.08
2011_Phys_FT	831	MC	05	1	775	0.6218							1.22
2011_Phys_FT	831	MC	06	1	775	−1.5464							0.94
2011_Phys_FT	831	MC	07	1	775	−0.8069							1.09
2011_Phys_FT	831	MC	08	1	775	−0.5354							0.95
2011_Phys_FT	831	MC	09	1	775	−0.6891							1.12
2011_Phys_FT	831	MC	10	1	775	−1.3074							1.04
2011_Phys_FT	831	MC	11	1	775	−0.4094							1.11
2011_Phys_FT	831	MC	12	1	775	−0.0113							1.03
2011_Phys_FT	831	CR	13	1	775	−0.6182							0.94
2011_Phys_FT	831	CR	14	1	775	−0.0548							1.10
2011_Phys_FT	831	CR	15	1	775	−0.7805							0.92
2011_Phys_FT	831	CR	16	1	775	−0.8401							0.86
2011_Phys_FT	831	CR	17	1	775	0.0887							0.81
2011_Phys_FT	831	CR	18	1	775	0.4016							0.76
2011_Phys_FT	831	CR	19	1	775	2.4258							0.87
2011_Phys_FT	831	CR	20	1	775	0.0324							0.81
2011_Phys_FT	832	MC	01	1	770	0.0356							1.01
2011_Phys_FT	832	MC	02	1	770	−0.6228							1.14
2011_Phys_FT	832	MC	03	1	770	−0.0446							1.06

Table 9. Partial Credit Model Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2011_Phys_FT	832	MC	04	1	770	−0.6356							1.12
2011_Phys_FT	832	MC	05	1	770	−1.5751							0.99
2011_Phys_FT	832	MC	06	1	770	−0.6677							1.23
2011_Phys_FT	832	MC	07	1	770	0.2725							1.15
2011_Phys_FT	832	MC	08	1	770	−0.8446							1.03
2011_Phys_FT	832	MC	09	1	770	−1.0082							0.90
2011_Phys_FT	832	MC	10	1	770	−0.3155							0.98
2011_Phys_FT	832	MC	11	1	770	0.3169							1.00
2011_Phys_FT	832	CR	12	1	770	−0.6036							0.92
2011_Phys_FT	832	CR	13	1	770	0.0603							1.00
2011_Phys_FT	832	CR	14	1	770	−0.9187							0.90
2011_Phys_FT	832	CR	15	1	770	−0.7718							0.93
2011_Phys_FT	832	CR	16	1	770	1.8163							0.93
2011_Phys_FT	832	CR	17	1	770	−0.0076							0.82
2011_Phys_FT	832	CR	18	1	770	−0.0815							0.81
2011_Phys_FT	832	CR	19	1	770	−0.1307							1.04
2011_Phys_FT	833	MC	01	1	791	−0.0664							1.01
2011_Phys_FT	833	MC	02	1	791	−0.9361							1.09
2011_Phys_FT	833	MC	03	1	791	0.3364							1.10
2011_Phys_FT	833	MC	04	1	791	−0.6527							1.03
2011_Phys_FT	833	MC	05	1	791	−0.3004							1.19
2011_Phys_FT	833	MC	06	1	791	−1.1933							1.04
2011_Phys_FT	833	MC	07	1	791	−0.0547							0.99
2011_Phys_FT	833	MC	08	1	791	0.0860							1.09

Table 9. Partial Credit Model Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2011_Phys_FT	833	MC	09	1	791	1.1262							1.37
2011_Phys_FT	833	MC	10	1	791	−2.5132							1.01
2011_Phys_FT	833	MC	11	1	791	−1.0621							1.07
2011_Phys_FT	833	CR	12	1	791	−2.0361							0.93
2011_Phys_FT	833	CR	13	1	791	0.6610							0.87
2011_Phys_FT	833	CR	14	1	791	−0.6405							0.84
2011_Phys_FT	833	CR	15	1	791	−1.0961							0.92
2011_Phys_FT	833	CR	16	1	791	0.2282							0.84
2011_Phys_FT	833	CR	17	1	791	0.8952							0.87
2011_Phys_FT	833	CR	18	1	791	−0.2476							0.79
2011_Phys_FT	833	CR	19	1	791	−0.1249							0.89
2011_Phys_FT	834	MC	01	1	774	−0.5579							1.03
2011_Phys_FT	834	MC	02	1	774	−0.8136							1.01
2011_Phys_FT	834	MC	03	1	774	−2.1102							1.02
2011_Phys_FT	834	MC	04	1	774	−1.1565							1.25
2011_Phys_FT	834	MC	05	1	774	−0.6354							0.94
2011_Phys_FT	834	MC	06	1	774	0.0457							1.05
2011_Phys_FT	834	MC	07	1	774	−0.2480							0.97
2011_Phys_FT	834	MC	08	1	774	−0.5965							1.00
2011_Phys_FT	834	MC	09	1	774	1.0159							1.02
2011_Phys_FT	834	MC	10	1	774	−0.9777							0.99
2011_Phys_FT	834	MC	11	1	774	−0.1293							1.10
2011_Phys_FT	834	CR	12	1	774	−1.1273							0.93
2011_Phys_FT	834	CR	13	1	774	−0.3295							0.93

Table 9. Partial Credit Model Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2011_Phys_FT	834	CR	14	1	774	−0.3107							0.90
2011_Phys_FT	834	CR	15	1	774	−1.0057							1.03
2011_Phys_FT	834	CR	16	1	774	1.1964							0.80
2011_Phys_FT	834	CR	17	1	774	−0.1980							0.80
2011_Phys_FT	834	CR	18	1	774	2.2049							0.99
2011_Phys_FT	834	CR	19	1	774	0.6084							1.18
2011_Phys_FT	835	MC	01	1	778	−0.9013							1.08
2011_Phys_FT	835	MC	02	1	778	−2.3271							1.05
2011_Phys_FT	835	MC	03	1	778	−0.1694							1.16
2011_Phys_FT	835	MC	04	1	778	−0.9430							1.03
2011_Phys_FT	835	MC	05	1	778	−0.1124							1.17
2011_Phys_FT	835	MC	06	1	778	−0.9500							1.18
2011_Phys_FT	835	MC	07	1	778	−0.5795							0.99
2011_Phys_FT	835	MC	08	1	778	0.9493							1.09
2011_Phys_FT	835	MC	09	1	778	−1.0492							1.01
2011_Phys_FT	835	MC	10	1	778	−0.0997							1.22
2011_Phys_FT	835	MC	11	1	778	0.6911							1.32
2011_Phys_FT	835	CR	12	1	778	−0.2455							0.99
2011_Phys_FT	835	CR	13	1	778	0.1035							0.96
2011_Phys_FT	835	CR	14	1	778	−0.7582							0.92
2011_Phys_FT	835	CR	15	1	778	0.3420							0.86
2011_Phys_FT	835	CR	16	1	778	−0.0617							0.74
2011_Phys_FT	835	CR	17	1	778	−0.5469							0.71
2011_Phys_FT	835	CR	18	1	778	0.8396							0.69

Table 9. Partial Credit Model Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2011_Phys_FT	835	CR	19	1	778	0.0017							0.78
2011_Phys_FT	836	MC	01	1	767	0.4487							0.98
2011_Phys_FT	836	MC	02	1	767	−0.1258							1.04
2011_Phys_FT	836	MC	03	1	767	−0.8736							0.99
2011_Phys_FT	836	MC	04	1	767	−0.2180							1.07
2011_Phys_FT	836	MC	05	1	767	0.0094							1.08
2011_Phys_FT	836	MC	06	1	767	−0.8200							0.95
2011_Phys_FT	836	MC	07	1	767	0.0402							1.09
2011_Phys_FT	836	MC	08	1	767	0.5599							1.09
2011_Phys_FT	836	MC	09	1	767	0.0217							1.10
2011_Phys_FT	836	MC	10	1	767	0.2703							1.10
2011_Phys_FT	836	MC	11	1	767	−0.3043							1.06
2011_Phys_FT	836	CR	12	1	767	0.1391							0.88
2011_Phys_FT	836	CR	13	1	767	−1.4065							0.82
2011_Phys_FT	836	CR	14	1	767	−0.4534							0.98
2011_Phys_FT	836	CR	15	1	767	−1.4552							1.03
2011_Phys_FT	836	CR	16	1	767	−1.7437							0.98
2011_Phys_FT	836	CR	17	1	767	0.4423							0.93
2011_Phys_FT	836	CR	18	1	767	−0.7736							0.81
2011_Phys_FT	836	CR	19	1	767	−0.9417							0.95
2011_Phys_FT	837	MC	01	1	752	−0.9616							1.08
2011_Phys_FT	837	MC	02	1	752	−1.6261							0.93
2011_Phys_FT	837	MC	03	1	752	0.2622							1.00
2011_Phys_FT	837	MC	04	1	752	−1.0783							1.06

Table 9. Partial Credit Model Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2011_Phys_FT	837	MC	05	1	752	−0.5911							1.22
2011_Phys_FT	837	MC	06	1	752	−1.3829							0.95
2011_Phys_FT	837	MC	07	1	752	−1.2460							1.08
2011_Phys_FT	837	MC	08	1	752	0.3403							1.33
2011_Phys_FT	837	MC	09	1	752	−1.0121							1.13
2011_Phys_FT	837	MC	10	1	752	0.6421							1.07
2011_Phys_FT	837	MC	11	1	752	−1.7008							0.91
2011_Phys_FT	837	CR	12	1	752	−1.9427							0.86
2011_Phys_FT	837	CR	13	1	752	1.0203							0.93
2011_Phys_FT	837	CR	14	1	752	−1.0561							0.88
2011_Phys_FT	837	CR	15	1	752	−1.2148							0.89
2011_Phys_FT	837	CR	16	1	752	1.0050							0.98
2011_Phys_FT	837	CR	17	1	752	0.1720							0.93
2011_Phys_FT	837	CR	18	1	752	−0.6845							0.87
2011_Phys_FT	837	CR	19	1	752	−0.1400							0.88
2011_Phys_FT	837	CR	20	1	752	−0.1527							0.92
2011_Phys_FT	838	MC	01	1	766	−0.6201							0.98
2011_Phys_FT	838	MC	02	1	766	0.7655							0.99
2011_Phys_FT	838	MC	03	1	766	0.6050							1.12
2011_Phys_FT	838	MC	04	1	766	0.6978							1.32
2011_Phys_FT	838	MC	05	1	766	−0.5388							1.01
2011_Phys_FT	838	MC	06	1	766	0.0137							1.07
2011_Phys_FT	838	MC	07	1	766	−0.0887							1.08
2011_Phys_FT	838	MC	08	1	766	0.2755							1.04

Table 9. Partial Credit Model Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2011_Phys_FT	838	MC	09	1	766	−0.3605							1.05
2011_Phys_FT	838	MC	10	1	766	−0.1248							1.02
2011_Phys_FT	838	MC	11	1	766	0.4251							1.12
2011_Phys_FT	838	CR	12	1	766	−0.7997							0.96
2011_Phys_FT	838	CR	13	1	766	−0.3727							0.99
2011_Phys_FT	838	CR	14	1	766	0.1165							0.96
2011_Phys_FT	838	CR	15										
2011_Phys_FT	838	CR	16	1	766	0.8416							0.89
2011_Phys_FT	838	CR	17	1	766	0.2018							0.86
2011_Phys_FT	838	CR	18	1	766	0.6978							0.82
2011_Phys_FT	838	CR	19	1	766	0.5206							0.80
2011_Phys_FT	838	CR	20	1	766	0.6711							0.92
2011_Phys_FT	839	MC	01	1	757	−0.5418							1.11
2011_Phys_FT	839	MC	02	1	757	−0.5817							1.27
2011_Phys_FT	839	MC	03	1	757	−1.1600							1.01
2011_Phys_FT	839	MC	04	1	757	−1.5763							1.01
2011_Phys_FT	839	MC	05	1	757	−0.3709							1.20
2011_Phys_FT	839	MC	06	1	757	−1.4723							0.97
2011_Phys_FT	839	MC	07	1	757	−1.8394							1.07
2011_Phys_FT	839	MC	08	1	757	−0.4823							1.14
2011_Phys_FT	839	MC	09	1	757	−1.2369							1.07
2011_Phys_FT	839	MC	10	1	757	1.3825							1.07
2011_Phys_FT	839	MC	11	1	757	0.2791							1.12
2011_Phys_FT	839	CR	12	1	757	−0.2216							0.84

Table 9. Partial Credit Model Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2011_Phys_FT	839	CR	13	1	757	−0.8336							0.84
2011_Phys_FT	839	CR	14	1	757	1.2433							1.04
2011_Phys_FT	839	CR	15	1	757	−0.8127							0.91
2011_Phys_FT	839	CR	16	1	757	0.4600							0.88
2011_Phys_FT	839	CR	17	1	757	−0.1181							0.82
2011_Phys_FT	839	CR	18	1	757	0.9087							0.90
2011_Phys_FT	839	CR	19	1	757	0.6537							0.83
2011_Phys_FT	839	CR	20	1	757	−0.6151							0.82
2011_Phys_FT	840	MC	01	1	756	−0.8433							0.97
2011_Phys_FT	840	MC	02	1	756	−1.5464							1.03
2011_Phys_FT	840	MC	03	1	756	0.1712							0.89
2011_Phys_FT	840	MC	04	1	756	0.9579							1.01
2011_Phys_FT	840	MC	05	1	756	0.3116							1.15
2011_Phys_FT	840	MC	06	1	756	−1.9821							0.95
2011_Phys_FT	840	MC	07	1	756	−1.1992							1.04
2011_Phys_FT	840	MC	08	1	756	−0.1373							1.15
2011_Phys_FT	840	MC	09	1	756	−1.2841							0.94
2011_Phys_FT	840	MC	10	1	756	−0.7878							1.14
2011_Phys_FT	840	MC	11	1	756	0.5313							1.24
2011_Phys_FT	840	MC	12	1	756	1.0051							1.12
2011_Phys_FT	840	CR	13	1	756	0.0787							0.91
2011_Phys_FT	840	CR	14	1	756	0.9269							0.87
2011_Phys_FT	840	CR	15	1	756	−0.2287							0.93
2011_Phys_FT	840	CR	16	1	756	1.2442							0.98

Table 9. Partial Credit Model Item Analysis *(continued)*

Test	Form	Type	Item	Max	N-Count	RID	S1	S2	S3	S4	S5	S6	INFIT
2011_Phys_FT	840	CR	17	1	756	0.6303							0.93
2011_Phys_FT	840	CR	18	1	756	0.5947							0.85
2011_Phys_FT	840	CR	19	1	756	0.9424							0.82
2011_Phys_FT	840	CR	20	1	756	−0.3005							0.96
2011_Phys_FT	841	MC	01	1	770	0.1300							1.27
2011_Phys_FT	841	MC	02	1	770	−0.7700							0.87
2011_Phys_FT	841	MC	03	1	770	−0.1200							1.18
2011_Phys_FT	841	MC	04	1	770	0.5600							0.92
2011_Phys_FT	841	MC	05	1	770	−2.1000							0.85
2011_Phys_FT	841	MC	06	1	770	−0.5000							1.07
2011_Phys_FT	841	MC	07	1	770	−0.0300							1.15
2011_Phys_FT	841	MC	08	1	770	−0.9600							1.06
2011_Phys_FT	841	MC	09	1	770	−1.9000							0.91
2011_Phys_FT	841	MC	10	1	770	0.0700							1.16
2011_Phys_FT	841	MC	11	1	770	0.6400							1.05
2011_Phys_FT	841	MC	12	1	770	−0.2400							1.01
2011_Phys_FT	841	CR	13	2	770	−0.1600	−0.0800	0.0800					0.93
2011_Phys_FT	841	CR	14	1	770	−0.1700							0.80
2011_Phys_FT	841	CR	15	2	770	−0.4900	−0.0300	0.0300					0.80
2011_Phys_FT	841	CR	16	2	770	0.4500	−1.0700	1.0700					0.84

Appendix C: DIF Statistics

Table 10. DIF Statistics

Form	Item	Item Type	MH Delta	MH Chi-Sq.	Effect Size	DIF Category	Favored Group
831	01	MC	0.23	0.36	0.03		
831	02	MC	−0.94	5.17	−0.12		
831	03	MC	0.02	0.00	−0.01		
831	04	MC	1.34	12.42	0.25	B	F
831	05	MC	−1.26	11.06	−0.25	B	M
831	06	MC	−0.45	0.93	−0.07		
831	07	MC	−0.59	2.31	−0.13		
831	08	MC	0.17	0.17	0.03		
831	09	MC	0.67	3.15	0.13		
831	10	MC	−0.08	0.04	−0.02		
831	11	MC	0.25	0.46	0.05		
831	12	MC	−0.12	0.09	−0.01		
831	13	CR		0.07	0.02		
831	14	CR		4.65	0.16		
831	15	CR		1.79	0.08		
831	16	CR		1.10	−0.06		
831	17	CR		0.35	−0.04		
831	18	CR		7.97	−0.17		
831	19	CR		0.54	0.04		
831	20	CR		0.63	0.05		
832	01	MC	−1.94	24.38	−0.33	C	M
832	02	MC	−1.04	8.02	−0.19	B	M
832	03	MC	−0.88	5.46	−0.16		
832	04	MC	−1.08	8.45	−0.21	B	M

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq.	Effect Size	DIF Category	Favored Group
832	05	MC	−0.15	0.11	−0.03		
832	06	MC	−0.22	0.37	−0.04		
832	07	MC	0.18	0.25	0.03		
832	08	MC	−0.45	1.30	−0.08		
832	09	MC	−0.53	1.49	−0.09		
832	10	MC	−0.34	0.76	−0.07		
832	11	MC	0.17	0.19	0.04		
832	12	CR		10.07	0.20	BB	F
832	13	CR		10.37	0.21	BB	F
832	14	CR		9.54	0.20	BB	F
832	15	CR		3.05	0.11		
832	16	CR		0.07	−0.02		
832	17	CR		10.26	0.18	BB	F
832	18	CR		8.04	0.17		
832	19	CR		0.40	0.05		
833	01	MC	−1.65	19.21	−0.29	C	M
833	02	MC	1.03	6.97	0.17	B	F
833	03	MC	−0.84	5.11	−0.16		
833	04	MC	−0.38	1.04	−0.05		
833	05	MC	0.30	0.71	0.07		
833	06	MC	−0.34	0.70	−0.06		
833	07	MC	−0.08	0.05	−0.03		
833	08	MC	−0.19	0.26	−0.04		
833	09	MC	−1.71	17.59	−0.30	C	M

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq.	Effect Size	DIF Category	Favored Group
833	10	MC	0.85	1.96	0.09		
833	11	MC	0.65	2.64	0.11		
833	12	CR		0.45	−0.04		
833	13	CR		2.34	0.11		
833	14	CR		0.21	0.02		
833	15	CR		0.32	0.03		
833	16	CR		0.04	0.00		
833	17	CR		2.75	0.09		
833	18	CR		2.75	0.09		
833	19	CR		7.26	0.17	BB	F
834	01	MC	0.68	3.15	0.11		
834	02	MC	0.26	0.43	0.05		
834	03	MC	0.58	1.23	0.06		
834	04	MC	−0.36	0.88	−0.06		
834	05	MC	0.17	0.18	0.04		
834	06	MC	0.58	2.32	0.10		
834	07	MC	0.91	5.74	0.14		
834	08	MC	−1.22	9.53	−0.21	B	M
834	09	MC	−0.53	1.54	−0.09		
834	10	MC	0.16	0.14	0.02		
834	11	MC	0.24	0.43	0.05		
834	12	CR		0.64	−0.05		
834	13	CR		0.58	−0.03		
834	14	CR		9.55	−0.19	BB	M

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq.	Effect Size	DIF Category	Favored Group
834	15	CR		3.04	0.11		
834	16	CR		0.04	0.02		
834	17	CR		2.71	0.09		
834	18	CR		3.86	−0.13		
834	19	CR		1.13	−0.07		
835	01	MC	−1.68	18.42	−0.29	C	M
835	02	MC	−1.90	12.27	−0.24	C	M
835	03	MC	−0.57	2.44	−0.11		
835	04	MC	0.84	4.30	0.14		
835	05	MC	0.04	0.01	0.00		
835	06	MC	0.11	0.09	0.03		
835	07	MC	0.19	0.21	0.02		
835	08	MC	−1.59	14.31	−0.25	C	M
835	09	MC	−1.07	6.83	−0.18	B	M
835	10	MC	−0.86	5.71	−0.17		
835	11	MC	0.81	4.85	0.17		
835	12	CR		2.29	0.09		
835	13	CR		13.79	0.23	BB	F
835	14	CR		1.88	0.08		
835	15	CR		5.83	0.14		
835	16	CR		2.44	0.09		
835	17	CR		0.01	0.00		
835	18	CR		0.05	0.01		
835	19	CR		5.26	0.12		

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq.	Effect Size	DIF Category	Favored Group
836	01	MC	−0.01	0.00	0.00		
836	02	MC	0.28	0.53	0.04		
836	03	MC	−0.37	0.83	−0.06		
836	04	MC	−0.53	2.06	−0.09		
836	05	MC	−0.83	4.90	−0.17		
836	06	MC	−0.42	1.05	−0.06		
836	07	MC	−0.40	1.15	−0.08		
836	08	MC	−0.13	0.11	−0.02		
836	09	MC	0.42	1.26	0.07		
836	10	MC	−0.12	0.11	−0.02		
836	11	MC	−0.93	5.96	−0.18		
836	12	CR		3.16	0.12		
836	13	CR		3.44	0.11		
836	14	CR		0.11	−0.01		
836	15	CR		7.43	0.19	BB	F
836	16	CR		2.40	0.11		
836	17	CR		0.77	0.05		
836	18	CR		0.13	−0.02		
836	19	CR		1.33	0.08		
837	01	MC	1.44	12.15	0.25	B	F
837	02	MC	−0.35	0.54	−0.06		
837	03	MC	−0.40	1.03	−0.05		
837	04	MC	−0.12	0.09	−0.04		
837	05	MC	−1.89	25.30	−0.36	C	M

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq.	Effect Size	DIF Category	Favored Group
837	06	MC	0.76	2.76	0.10		
837	07	MC	−0.54	1.76	−0.11		
837	08	MC	−0.52	2.00	−0.11		
837	09	MC	−0.84	4.50	−0.14		
837	10	MC	0.37	0.86	0.06		
837	11	MC	1.54	9.04	0.22	C	F
837	12	CR		0.85	0.05		
837	13	CR		2.56	0.10		
837	14	CR		3.80	0.11		
837	15	CR		2.41	0.12		
837	16	CR		1.37	−0.08		
837	17	CR		0.01	0.00		
837	18	CR		1.14	0.07		
837	19	CR		4.04	−0.13		
837	20	CR		1.14	0.07		
838	01	MC	0.17	0.18	0.03		
838	02	MC	−0.36	0.76	−0.07		
838	03	MC	0.24	0.41	0.03		
838	04	MC	−0.01	0.00	−0.01		
838	05	MC	−1.01	6.97	−0.17	B	M
838	06	MC	0.05	0.02	0.02		
838	07	MC	−0.97	7.24	−0.18		
838	08	MC	−0.57	2.26	−0.12		
838	09	MC	−1.74	21.01	−0.32	C	M

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq.	Effect Size	DIF Category	Favored Group
838	10	MC	−1.58	16.78	−0.28	C	M
838	11	MC	0.24	0.43	0.05		
838	12	CR		20.33	0.31	CC	F
838	13	CR		4.97	0.16		
838	14	CR		5.31	−0.15		
838	16	CR		16.01	0.24	BB	F
838	17	CR		6.41	0.15		
838	18	CR		16.50	0.25	BB	F
838	19	CR		9.28	0.17	BB	F
838	20	CR		1.58	−0.07		
839	01	MC	0.54	2.06	0.11		
839	02	MC	−0.26	0.51	−0.04		
839	03	MC	−0.80	3.56	−0.14		
839	04	MC	−0.16	0.11	−0.01		
839	05	MC	−0.11	0.09	−0.02		
839	06	MC	0.25	0.30	0.03		
839	07	MC	−1.43	8.90	−0.19	B	M
839	08	MC	−1.03	7.79	−0.20	B	M
839	09	MC	−0.44	1.12	−0.08		
839	10	MC	0.20	0.20	0.03		
839	11	MC	0.00	0.00	0.00		
839	12	CR		1.98	0.08		
839	13	CR		6.21	0.14		
839	14	CR		0.32	0.04		

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq.	Effect Size	DIF Category	Favored Group
839	15	CR		4.77	0.14		
839	16	CR		1.52	0.07		
839	17	CR		0.13	−0.02		
839	18	CR		5.41	−0.15		
839	19	CR		5.79	0.13		
839	20	CR		0.26	0.03		
840	01	MC	0.73	3.08	0.10		
840	02	MC	0.33	0.51	0.03		
840	03	MC	0.46	1.15	0.07		
840	04	MC	−1.15	6.58	−0.16	B	M
840	05	MC	−0.41	1.18	−0.07		
840	06	MC	−0.21	0.16	−0.02		
840	07	MC	−0.83	3.82	−0.13		
840	08	MC	0.13	0.13	0.01		
840	09	MC	−1.28	8.15	−0.21	B	M
840	10	MC	1.49	14.23	0.25	B	F
840	11	MC	0.08	0.04	0.02		
840	12	MC	−1.18	7.77	−0.16	B	M
840	13	CR		0.00	0.00		
840	14	CR		0.12	0.01		
840	15	CR		4.44	0.14		
840	16	CR		0.66	0.07		
840	17	CR		0.41	0.03		
840	18	CR		0.00	0.02		

Table 10. DIF Statistics *(continued)*

Form	Item	Item Type	MH Delta	MH Chi-Sq.	Effect Size	DIF Category	Favored Group
840	19	CR		0.43	0.05		
840	20	CR		1.44	−0.09		
841	01	MC	0.15	0.17	0.02		
841	02	MC	0.44	1.11	0.06		
841	03	MC	−0.37	1.05	−0.07		
841	04	MC	−1.12	7.11	−0.19	B	M
841	05	MC	1.02	3.19	0.14		
841	06	MC	0.26	0.48	0.05		
841	07	MC	−0.74	4.18	−0.13		
841	08	MC	1.34	11.27	0.25	B	F
841	09	MC	−1.08	4.36	−0.13	B	M
841	10	MC	−1.87	26.01	−0.38	C	M
841	11	MC	−0.27	0.47	−0.06		
841	12	MC	−0.16	0.17	−0.03		
841	13	CR		2.89	0.09		
841	14	CR		0.53	0.04		
841	15	CR		3.43	0.09		
841	16	CR		1.68	0.08		

*DIF Category meanings: A/AA=negligible, B/BB=moderate, C/CC=large

Appendix D: Operational Test Maps

Table 11. Operational Test Map for January 2011

Position	Item Type	Max Points	Weight	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
1	MC	1	1	0.28	0.29	0.90						
2	MC	1	1	0.69	0.41	−1.10						
3	MC	1	1	0.38	0.37	0.40						
4	MC	1	1	0.52	0.32	−0.27						
5	MC	1	1	0.59	0.37	−0.65						
6	MC	1	1	0.76	0.36	−1.47						
7	MC	1	1	0.87	0.32	−2.25						
8	MC	1	1	0.53	0.43	−0.35						
9	MC	1	1	0.37	0.44	0.43						
10	MC	1	1	0.42	0.43	0.19						
11	MC	1	1	0.59	0.33	−0.57						
12	MC	1	1	0.68	0.39	−1.04						
13	MC	1	1	0.66	0.42	−0.92						
14	MC	1	1	0.63	0.40	−0.80						
15	MC	1	1	0.67	0.42	−0.96						
16	MC	1	1	0.85	0.37	−2.17						
17	MC	1	1	0.36	0.36	0.50						
18	MC	1	1	0.66	0.45	−0.97						
19	MC	1	1	0.46	0.41	0.02						
20	MC	1	1	0.38	0.28	0.42						
21	MC	1	1	0.40	0.03	0.31						
22	MC	1	1	0.50	0.38	−0.17						
23	MC	1	1	0.68	0.33	−1.06						

Table 11. Operational Test Map for January 2011 *(continued)*

Position	Item Type	Max Points	Weight	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
24	MC	1	1	0.66	0.37	−0.90						
25	MC	1	1	0.81	0.46	−1.85						
26	MC	1	1	0.88	0.34	−2.52						
27	MC	1	1	0.66	0.39	−0.94						
28	MC	1	1	0.47	0.33	−0.04						
29	MC	1	1	0.48	0.18	−0.10						
30	MC	1	1	0.78	0.39	−1.63						
31	MC	1	1	0.71	0.38	−1.19						
32	MC	1	1	0.29	0.15	0.85						
33	MC	1	1	0.42	0.28	0.19						
34	MC	1	1	0.64	0.31	−0.82						
35	MC	1	1	0.54	0.38	−0.37						
36	MC	1	1	0.34	0.30	0.57						
37	MC	1	1	0.45	0.42	0.08						
38	MC	1	1	0.72	0.41	−1.29						
39	MC	1	1	0.42	0.31	0.22						
40	MC	1	1	0.63	0.25	−0.80						
41	MC	1	1	0.31	0.33	0.74						
42	MC	1	1	0.36	0.32	0.49						
43	MC	1	1	0.72	0.31	−1.26						
44	MC	1	1	0.54	0.48	−0.34						
45	MC	1	1	0.71	0.32	−1.17						
46	MC	1	1	0.74	0.37	−1.35						

Table 11. Operational Test Map for January 2011 *(continued)*

Position	Item Type	Max Points	Weight	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
47	MC	1	1	0.74	0.35	−1.35						
48	MC	1	1	0.36	0.40	0.52						
49	MC	1	1	0.17	0.38	1.65						
50	MC	1	1	0.29	0.26	0.86						
51	CR	1	1	0.84	0.38	−2.04						
52	CR	1	1	0.67	0.51	−0.98						
53	CR	2	1	1.35	0.62	−0.84	0.21	−0.21				
54	CR	1	1	0.34	0.23	0.65						
55	CR	1	1	0.37	0.45	0.47						
56	CR	1	1	0.15	0.41	1.80						
57	CR	2	1	0.76	0.61	0.27	0.22	−0.22				
58	CR	1	1	0.38	0.49	0.43						
59	CR	2	1	1.43	0.52	−1.14	−0.31	0.31				
60	CR	2	1	1.11	0.63	−0.43	−0.53	0.53				
61	CR	1	1	0.22	0.47	1.28						
62	CR	1	1	0.72	0.39	−1.24						
63	CR	1	1	0.52	0.45	−0.28						
64	CR	2	1	0.46	0.57	1.08	−0.29	0.29				
65	CR	1	1	0.21	0.56	1.36						
66	CR	2	1	1.43	0.57	−1.13	−0.44	0.44				
67	CR	2	1	1.17	0.64	−0.50	−0.19	0.19				
68	CR	1	1	0.64	0.54	−0.84						
69	CR	1	1	0.24	0.39	1.18						

Table 11. Operational Test Map for January 2011 *(continued)*

Position	Item Type	Max Points	Weight	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
70	CR	1	1	0.52	0.46	−0.26						
71	CR	2	1	0.59	0.66	0.74	−0.29	0.29				
72	CR	1	1	0.38	0.57	0.43						
73	CR	1	1	0.69	0.46	−1.06						
74	CR	2	1	0.40	0.62	1.28	−0.43	0.43				
75	CR	2	1	0.45	0.61	0.96	0.10	−0.10				

Table 12. Operational Test Map for June 2011

Position	Item Type	Max Points	Weight	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
1	MC	1	1	0.55	0.42	−0.49						
2	MC	1	1	0.51	0.37	−0.31						
3	MC	1	1	0.88	0.35	−2.72						
4	MC	1	1	0.68	0.43	−1.10						
5	MC	1	1	0.32	0.23	0.63						
6	MC	1	1	0.60	0.31	−0.75						
7	MC	1	1	0.82	0.32	−2.00						
8	MC	1	1	0.47	0.34	−0.11						
9	MC	1	1	0.60	0.31	−0.72						
10	MC	1	1	0.65	0.44	−0.99						
11	MC	1	1	0.26	0.38	1.09						
12	MC	1	1	0.38	0.17	0.34						
13	MC	1	1	0.42	0.28	0.11						
14	MC	1	1	0.40	0.46	0.29						
15	MC	1	1	0.65	0.37	−0.94						
16	MC	1	1	0.30	0.28	0.76						
17	MC	1	1	0.79	0.26	−1.77						
18	MC	1	1	0.58	0.44	−0.66						
19	MC	1	1	0.49	0.41	−0.19						
20	MC	1	1	0.75	0.40	−1.62						
21	MC	1	1	0.34	0.27	0.56						
22	MC	1	1	0.63	0.44	−0.84						
23	MC	1	1	0.49	0.48	−0.20						

Table 12. Operational Test Map for June 2011 *(continued)*

Position	Item Type	Max Points	Weight	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
24	MC	1	1	0.82	0.31	−2.00						
25	MC	1	1	0.78	0.35	−1.79						
26	MC	1	1	0.54	0.29	−0.41						
27	MC	1	1	0.58	0.31	−0.61						
28	MC	1	1	0.43	0.28	0.10						
29	MC	1	1	0.63	0.36	−0.89						
30	MC	1	1	0.39	0.28	0.26						
31	MC	1	1	0.86	0.43	−2.28						
32	MC	1	1	0.41	0.33	0.20						
33	MC	1	1	0.65	0.30	−1.01						
34	MC	1	1	0.31	0.11	0.67						
35	MC	1	1	0.23	0.26	1.27						
36	MC	1	1	0.68	0.43	−1.10						
37	MC	1	1	0.29	0.43	0.79						
38	MC	1	1	0.53	0.32	−0.40						
39	MC	1	1	0.68	0.40	−1.11						
40	MC	1	1	0.52	0.41	−0.35						
41	MC	1	1	0.45	0.34	−0.03						
42	MC	1	1	0.36	0.35	0.42						
43	MC	1	1	0.61	0.30	−0.81						
44	MC	1	1	0.42	0.35	0.14						
45	MC	1	1	0.79	0.27	−1.76						
46	MC	1	1	0.59	0.37	−0.70						

Table 12. Operational Test Map for June 2011 *(continued)*

Position	Item Type	Max Points	Weight	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
47	MC	1	1	0.59	0.43	−0.68						
48	MC	1	1	0.62	0.37	−0.81						
49	MC	1	1	0.43	0.31	0.09						
50	MC	1	1	0.37	0.26	0.37						
51	CR	1	1	0.69	0.45	−1.20						
52	CR	1	1	0.76	0.47	−1.63						
53	CR	1	1	0.64	0.60	−0.98						
54	CR	1	1	0.50	0.62	−0.24						
55	CR	1	1	0.63	0.64	−0.91						
56	CR	1	1	0.56	0.46	−0.53						
57	CR	1	1	0.55	0.45	−0.46						
58	CR	1	1	0.62	0.56	−0.86						
59	CR	1	1	0.32	0.50	0.66						
60	CR	1	1	0.40	0.56	0.26						
61	CR	1	1	0.55	0.55	−0.47						
62	CR	1	1	0.33	0.40	0.57						
63	CR	1	1	0.31	0.53	0.73						
64	CR	1	1	0.43	0.55	0.10						
65	CR	1	1	0.20	0.46	1.38						
66	CR	1	1	0.60	0.51	−0.71						
67	CR	1	1	0.32	0.54	0.61						
68	CR	1	1	0.47	0.57	−0.12						
69	CR	1	1	0.32	0.52	0.60						

Table 12. Operational Test Map for June 2011 *(continued)*

Position	Item Type	Max Points	Weight	Mean	Point-Biserial	Rasch	S1	S2	S3	S4	S5	S6
70	CR	1	1	0.30	0.66	0.86						
71	CR	1	1	0.50	0.72	−0.22						
72	CR	1	1	0.30	0.63	0.85						
73	CR	1	1	0.41	0.65	0.22						
74	CR	1	1	0.52	0.57	−0.32						
75	CR	1	1	0.28	0.56	0.91						
76	CR	1	1	0.24	0.59	1.10						
77	CR	1	1	0.55	0.45	−0.48						
78	CR	1	1	0.33	0.58	0.60						
79	CR	1	1	0.54	0.60	−0.43						
80	CR	1	1	0.38	0.63	0.33						
81	CR	1	1	0.44	0.66	0.04						
82	CR	1	1	0.40	0.52	0.20						
83	CR	1	1	0.46	0.54	−0.06						
84	CR	1	1	0.34	0.53	0.54						
85	CR	1	1	0.06	0.26	2.86						

Appendix E: Scoring Tables

Table 13. Scoring Table for January 2011

Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score
0	−6.291	0.000		23	−1.407	35.062		46	−0.053	62.876		69	1.428	86.128
1	−5.070	1.948		24	−1.339	36.365		47	0.003	63.995		70	1.514	86.991
2	−4.352	3.854		25	−1.274	37.677		48	0.059	65.099		71	1.604	87.929
3	−3.922	5.547		26	−1.209	38.996		49	0.116	66.208		72	1.699	88.851
4	−3.610	7.296		27	−1.146	40.280		50	0.173	67.287		73	1.800	89.669
5	−3.362	8.931		28	−1.083	41.512		51	0.230	68.321		74	1.907	90.552
6	−3.156	10.562		29	−1.022	42.757		52	0.288	69.430		75	2.023	91.438
7	−2.978	12.192		30	−0.962	44.006		53	0.346	70.464		76	2.148	92.308
8	−2.820	13.809		31	−0.902	45.262		54	0.405	71.516		77	2.285	93.170
9	−2.679	15.339		32	−0.843	46.526		55	0.464	72.522		78	2.438	94.027
10	−2.550	16.853		33	−0.785	47.724		56	0.524	73.580		79	2.612	94.879
11	−2.431	18.378		34	−0.727	48.966		57	0.585	74.573		80	2.814	95.715
12	−2.321	19.831		35	−0.670	50.178		58	0.647	75.536		81	3.056	96.593
13	−2.217	21.252		36	−0.613	51.359		59	0.710	76.576		82	3.363	97.418
14	−2.119	22.679		37	−0.556	52.553		60	0.774	77.543		83	3.788	98.206
15	−2.027	24.116		38	−0.500	53.745		61	0.839	78.558		84	4.500	99.231
16	−1.939	25.555		39	−0.444	54.941		62	0.906	79.510		85	5.717	100.000
17	−1.854	26.954		40	−0.388	56.060		63	0.974	80.509				
18	−1.773	28.317		41	−0.332	57.237		64	1.044	81.415				
19	−1.695	29.679		42	−0.277	58.396		65	1.116	82.344				
20	−1.620	31.043		43	−0.221	59.554		66	1.190	83.317				
21	−1.547	32.420		44	−0.165	60.659		67	1.266	84.212				
22	−1.476	33.766		45	−0.109	61.775		68	1.345	85.181				

Table 14. Scoring Table for June 2011

Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score		Raw Score	Ability	Scale Score
0	−6.311	0.000		23	−1.393	35.330		46	−0.040	63.150		69	1.448	86.334
1	−5.088	1.901		24	−1.325	36.656		47	0.016	64.257		70	1.537	87.229
2	−4.368	3.793		25	−1.258	37.991		48	0.072	65.350		71	1.629	88.175
3	−3.935	5.470		26	−1.193	39.328		49	0.128	66.449		72	1.727	89.087
4	−3.621	7.217		27	−1.129	40.607		50	0.184	67.490		73	1.831	89.926
5	−3.372	8.859		28	−1.066	41.855		51	0.241	68.537		74	1.942	90.840
6	−3.164	10.497		29	−1.005	43.118		52	0.299	69.640		75	2.062	91.733
7	−2.984	12.133		30	−0.944	44.373		53	0.356	70.639		76	2.192	92.602
8	−2.825	13.759		31	−0.884	45.648		54	0.415	71.705		77	2.335	93.460
9	−2.682	15.304		32	−0.825	46.914		55	0.474	72.685		78	2.494	94.305
10	−2.552	16.834		33	−0.767	48.088		56	0.534	73.750		79	2.675	95.139
11	−2.431	18.376		34	−0.709	49.366		57	0.595	74.725		80	2.884	95.979
12	−2.319	19.849		35	−0.652	50.554		58	0.657	75.699		81	3.136	96.878
13	−2.214	21.291		36	−0.595	51.737		59	0.720	76.724		82	3.454	97.632
14	−2.115	22.741		37	−0.538	52.930		60	0.784	77.706		83	3.891	98.352
15	−2.021	24.204		38	−0.482	54.123		61	0.850	78.708		84	4.620	99.415
16	−1.932	25.668		39	−0.427	55.277		62	0.917	79.680		85	5.850	100.000
17	−1.846	27.085		40	−0.371	56.423		63	0.986	80.665				
18	−1.764	28.475		41	−0.316	57.565		64	1.057	81.577				
19	−1.685	29.861		42	−0.261	58.747		65	1.130	82.535				
20	−1.609	31.254		43	−0.205	59.857		66	1.205	83.492				
21	−1.535	32.656		44	−0.150	60.958		67	1.283	84.419				
22	−1.463	34.011		45	−0.095	62.062		68	1.364	85.400				